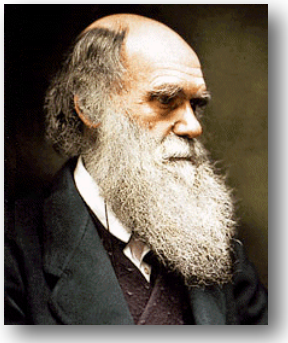
Positive selection on non-coding sequences during primate evolution: from genome to nucleotide

Gregory Wray

Department of Biology
and
Institute for Genome Sciences & Policy

Duke University
Durham, NC, USA

The great apes and an outgroup

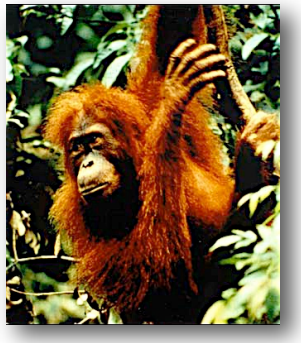


Homo sapiens

Pan troglodytes

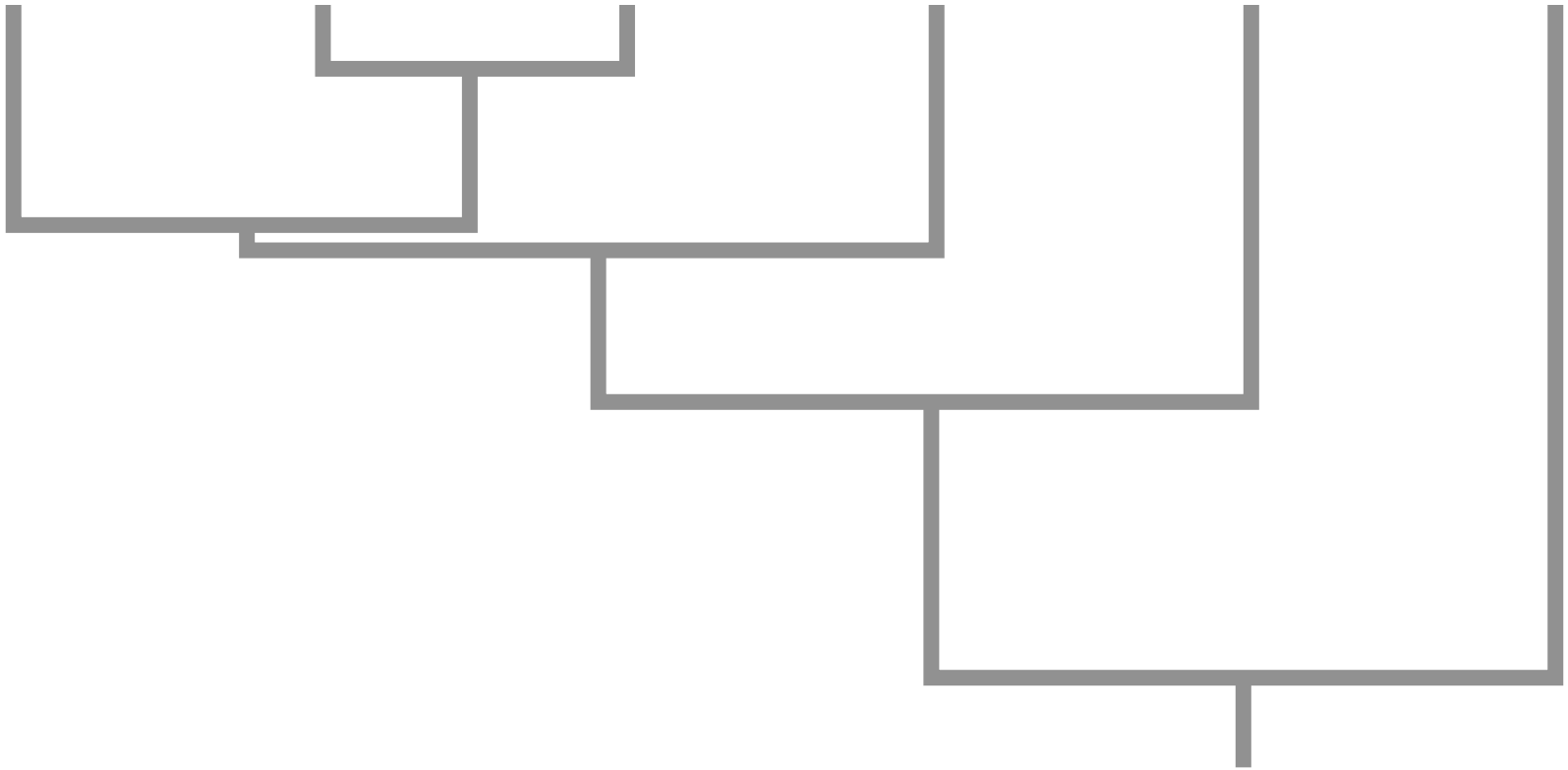
Pan paniscus

Gorilla gorilla



Pongo pygmaeus

Macaca mulatta



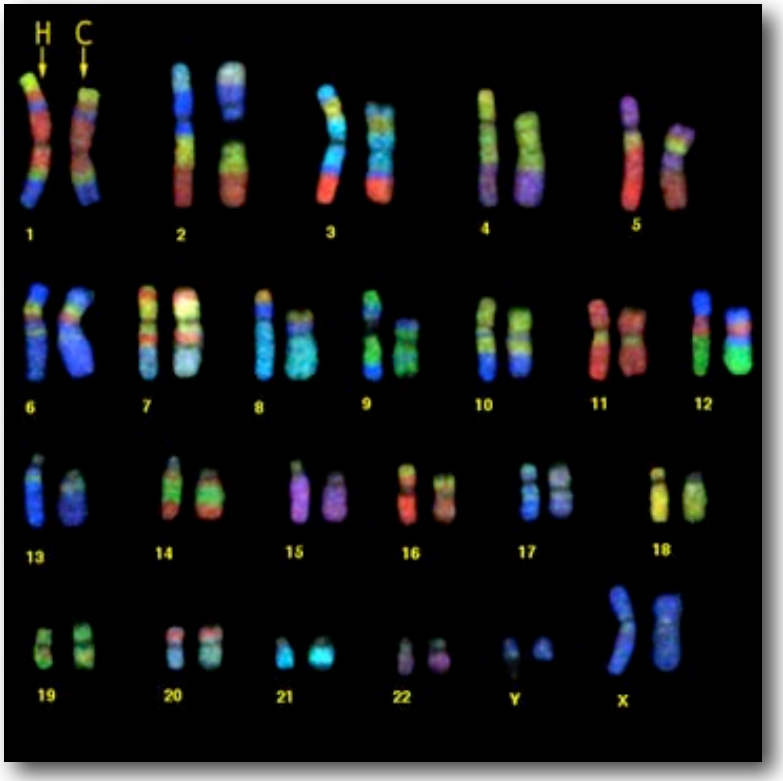
An apparent paradox

Genotype: quite similar

Very similar gene complement

Many identical proteins

~98.8% sequence identity



Phenotype: many differences

Cognition and behavior

Morphology

Physiology and reproduction

Disease susceptibility

Michael Neugebauer

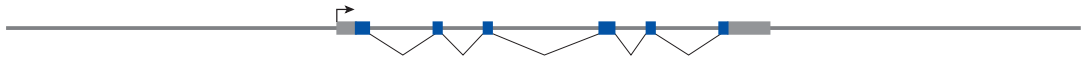
King and Wilson's proposal

Phenotypic differences between humans and chimpanzees are due primarily to changes in the regulation of gene expression

King, M-C and A Wilson (1975) *Science* 188:107

The problem: finding a needle in a field of haystacks

GENOME



GENE

AGCCTAGGTAGCCTATTACGCGACCTAGGTACTACTTCGGGACTACCGAGCCTAGGTAGCCTAT

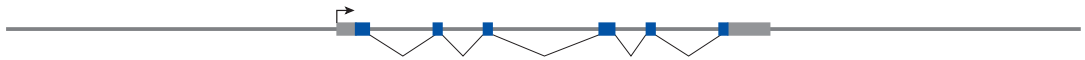
MUTATION

The problem: finding a needle in a field of haystacks

~24,000 genes

1 gene

GENOME



GENE

AGCCTAGGTAGCCTATTACGCGACCTAGGTACTACTTCGGGACTACCGAGCCTAGGTAGCCTAT

MUTATION

The problem: finding a needle in a field of haystacks

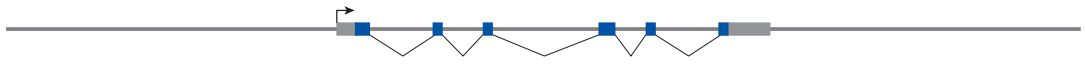
~24,000 genes

1 gene

10,000's of nucleotides

1 nucleotide

GENOME



GENE

AGCCTAGGTAGCCTATTACGCGACCTAGGTACTACTTCGGGACTACCGAGCCTAGGTAGCCTAT

MUTATION

Finding the needles

Candidate gene approach:

Think of genes that seem plausibly associated with trait of interest

Figure out a way to test for functional & phenotypic consequences

Often works, but limited to a few, well-studied genes

Finding the needles

Candidate gene approach:

Think of genes that seem plausibly associated with trait of interest

Figure out a way to test for functional & phenotypic consequences

Often works, but limited to a few, well-studied genes

Open-ended screen:

Identify genes that show evidence of recent positive selection

Figure out a way to test for functional & phenotypic consequences

Unbiased, considers all genes (even if no functional data)

screening for more candidate genes

testing for functional differences

connecting the dots: 3 case studies

screening for more candidate genes

testing for functional differences

connecting the dots: 3 case studies

A scan for positive selection on non-coding sequences

Goals: unbiased, genome-wide, statistically robust

Three-genome comparison

Human, chimpanzee, macaque

Scan non-coding regions near as many genes as possible

Test for selection using a random-effects likelihood approach

Use introns as local neutral proxies

Estimate substitution rates under different models

Use likelihood ratio to test for significance

Haygood et al. (2007) *Nature Genetics* 39:1140-1144.

Testing for positive selection in non-coding sequences

$$\zeta = \frac{\text{query}}{\text{reference}} = \frac{\text{red}}{\text{green}}$$

$\zeta < 1$

negative selection

functional constraint

$\zeta = 1$

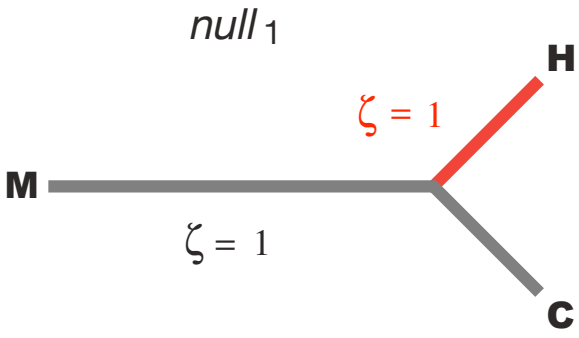
drift

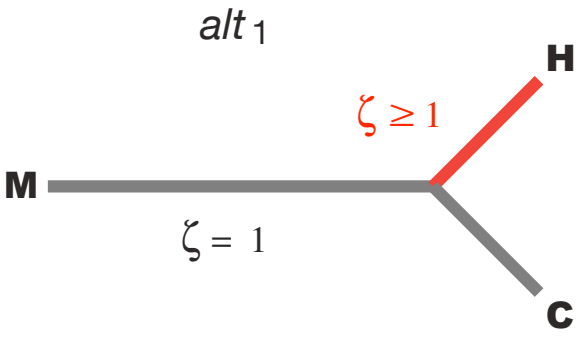
$\zeta > 1$

positive selection

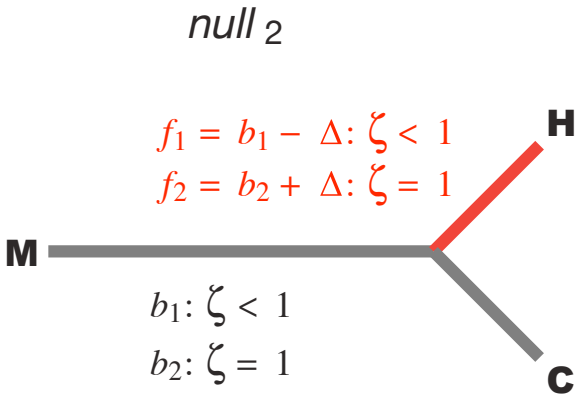
adaptation

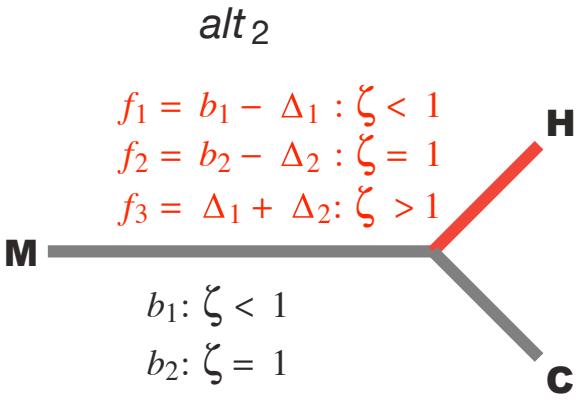
Likelihood-based tests for positive selection





fewer assumptions,
unrealistic
less sensitive

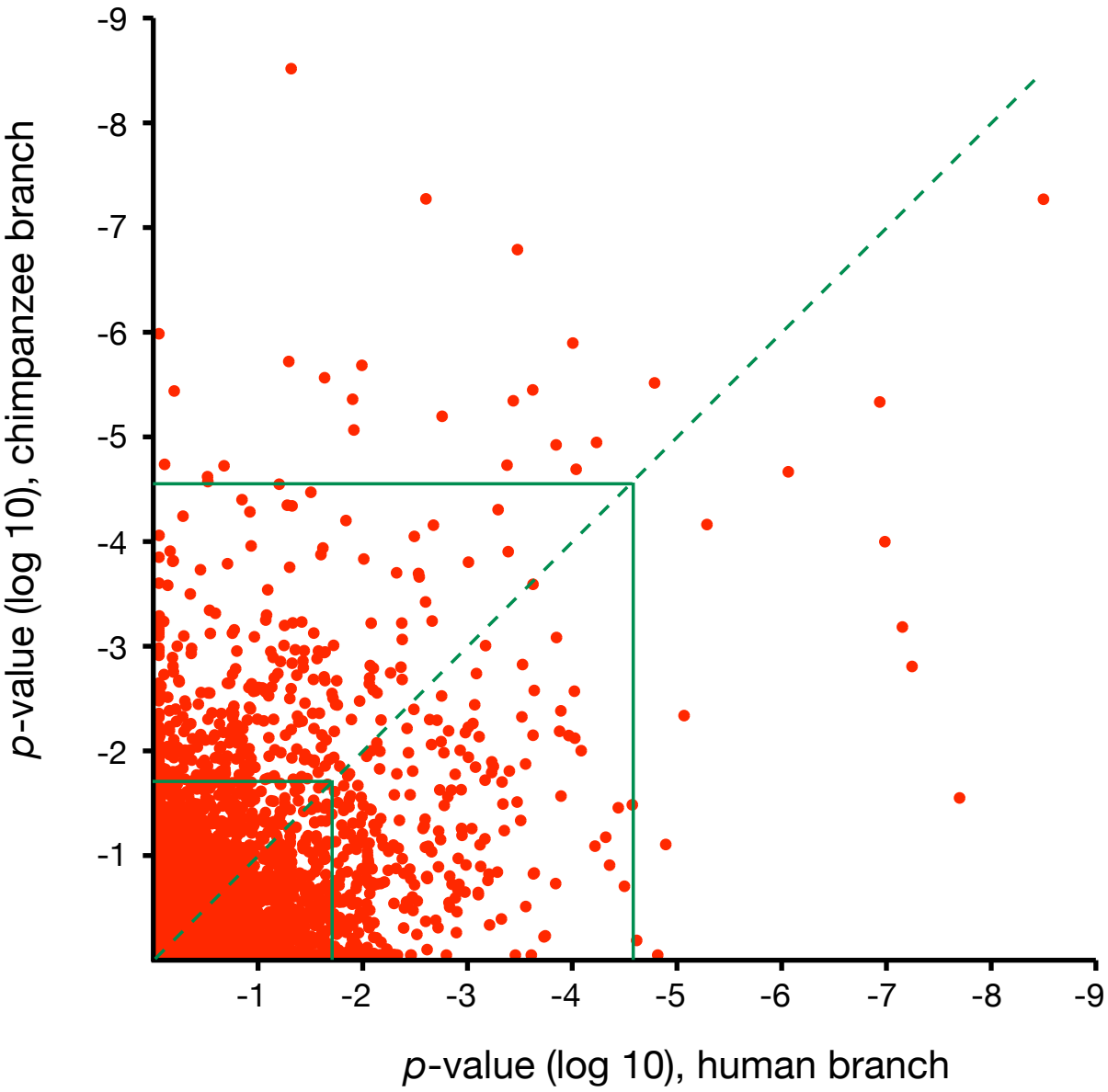




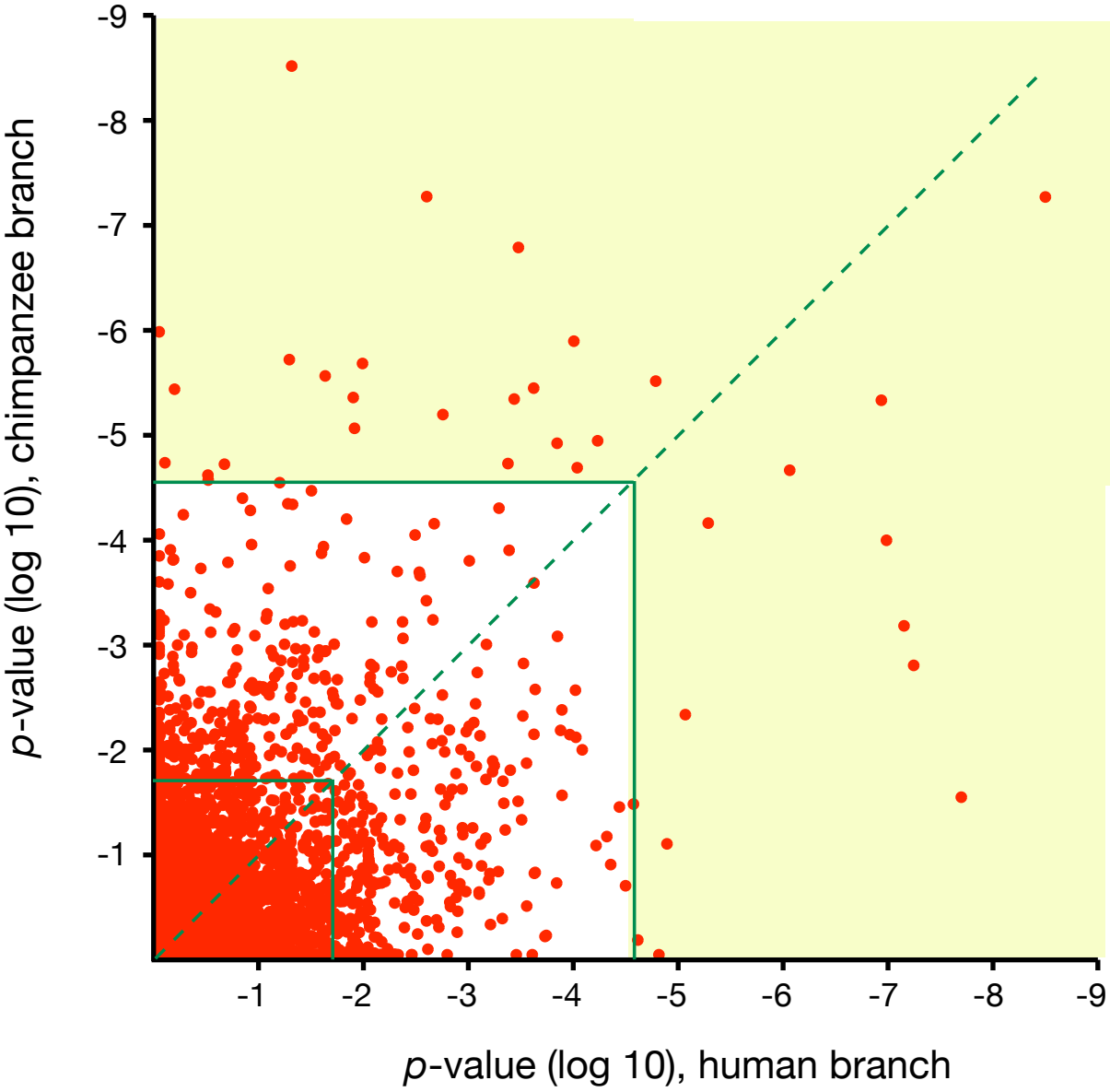
more realistic,
more sensitive,
harder to estimate

Conceptual approach modified from: Wong and Nielsen (2004) *Genetics* 167:949-958

Distribution of *p*-values on chimp and human branches



Distribution of *p*-values on chimp and human branches



No signal of adaptation in neural function from coding regions

Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios

Andrew G. Clark,[1] Stephen Glanowski,[3] Rasmus Nielsen,[2] Paul D. Thomas,[4] Anish Kejariwal,[4] Melissa A. Todd,[2] David M. Tanenbaum,[5] Daniel Civello,[6] Fu Lu,[5] Brian Murphy,[3] Steve Ferriera,[3] Gary Wang,[3] Xianqgun Zheng,[5] Thomas J. White,[6] John J. Sninsky,[6] Mark D. Adams,[5]* Michele Cargill[6]†

Even though human and chimpanzee gene sequences are nearly 99% identical, sequence comparisons can nevertheless be highly informative in identifying biologically important changes that have occurred since our ancestral lineages diverged. We analyzed alignments of 7645 chimpanzee gene sequences to their human and mouse orthologs. These three-species sequence alignments allowed us to identify genes undergoing natural selection along the human and chimp lineage by fitting models that include parameters specifying rates of synonymous and nonsynonymous nucleotide substitution. This evolutionary approach revealed an informative set of genes with significantly different patterns of substitution on the human lineage compared with the chimpanzee and mouse lineages. Partitions of genes into inferred biological classes identified accelerated evolution in several functional classes, including olfaction and nuclear transport. In addition to suggesting adaptive physiological differences between chimps and humans, human-accelerated genes are significantly more likely to underlie major known Mendelian disorders.

Although the human genome project will allow us to compare our genome to that of other primates and discover features that are uniquely human, there is no guarantee that such features are responsible for any of our unique biological attributes. To identify genes and biological processes that have been most altered by our recent evolutionary divergence from other primates, we need to fit the data to models of sequence divergence that allow us to distinguish between divergence caused by random drift and divergence driven by natural selection. Early observations of unexpectedly low levels of protein divergence between humans and chimpanzees led to the hypothesis that most of the evolutionary changes must have occurred at the level of gene regulation (*1*). Recently, much more extensive efforts at DNA sequencing in nonhuman primates has confirmed the very close evolutionary relationship between humans and chimps (*2*), with an average nucleotide divergence of just 1.2% (*3–5*). The role of protein divergence in causing morphological, physiological, and behavioral differences between these two species, however, remains unknown.

Here we apply evolutionary tests to identify genes and pathways from a new collection of more than 200,000 chimpanzee exonic sequences that show patterns of divergence consistent with natural selection along the human and chimpanzee lineages.

To construct the human-chimp-mouse alignments, we sequenced PCR amplifications using primers designed to essentially all human exons from one male chimpanzee, resulting in more than 20,000 human-chimp gene alignments spanning 18.5 Mb (*6–8*). To identify changes that are specific to the divergence in the human lineage, we compared the human-chimp aligned genes to their mouse ortholog. Inference of orthology involved a combination of reciprocal best matches and syntenic evidence between human and mouse gene annotations (*9*, *10*). This genome-wide set of orthologs underwent a series of filtering steps to remove ambiguities, orthologs with little sequence data, and genes with suspect annotation (*6*). The filtered ortholog set was compared to

[1]Molecular Biology and Genetics, [2]Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA. [3]Applied Biosystems, 45 West Gude Drive, Rockville, MD 20850, USA. [4]Protein Informatics, Celera Genomics, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [5]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [6]Celera Diagnostics, 1401 Harbor Bay Parkway, Alameda, CA 94502, USA.

*Present address: Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA.

†To whom correspondence should be addressed. E-mail: michele_cargill@celeradiagnostics.com

Clark et al. (2003) *Science* **302**:1960-1963.

A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees

Rasmus Nielsen[1,2*], Carlos Bustamante[1], Andrew G. Clark[3], Stephen Glanowski[4], Timothy B. Sackton[3], Melissa J. Hubisz[1], Adi Fledel-Alon[1], David M. Tanenbaum[5], Daniel Civello[6], Thomas J. White[6], John J. Sninsky[6], Mark D. Adams[5¤], Michele Cargill[6]

1 Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, **2** Center for Bioinformatics, University of Copenhagen, Denmark, **3** Molecular Biology and Genetics, Cornell University, Ithaca, New York, United States of America, **4** Applied Biosystems, Rockville, Maryland, United States of America, **5** Celera Genomics, Rockville, Maryland, United States of America, **6** Celera Diagnostics, Alameda, California, United States of America

Since the divergence of humans and chimpanzees about 5 million years ago, these species have undergone a remarkable evolution with drastic divergence in anatomy and cognitive abilities. At the molecular level, despite the small overall magnitude of DNA sequence divergence, we might expect such evolutionary changes to leave a noticeable signature throughout the genome. We here compare 13,731 annotated genes from humans to their chimpanzee orthologs to identify genes that show evidence of positive selection. Many of the genes that present a signature of positive selection tend to be involved in sensory perception or immune defenses. However, the group of genes that show the strongest evidence for positive selection also includes a surprising number of genes involved in tumor suppression and apoptosis, and of genes involved in spermatogenesis. We hypothesize that positive selection in some of these genes may be driven by genomic conflict due to apoptosis during spermatogenesis. Genes with maximal expression in the brain show little or no evidence for positive selection, while genes with maximal expression in the testis tend to be enriched with positively selected genes. Genes on the X chromosome also tend to show an elevated tendency for positive selection. We also present polymorphism data from 20 Caucasian Americans and 19 African Americans for the 50 annotated genes showing the strongest evidence for positive selection. The polymorphism analysis further supports the presence of positive selection in these genes by showing an excess of high-frequency derived nonsynonymous mutations.

Citation: Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3(6): e170.

Introduction

Genes, or regions of the genome, that have been affected by natural selection may show an excess of functionally important molecular changes, beyond what would be expected in the absence of selection. Genomic regions with such an excess of changes are said to have experienced positive selection, i.e., selection in favor of new genetic variants. The most common statistical technique for detecting positive selection takes advantage of the fact that mutations in coding regions of genes come in two classes: nonsynonymous mutations that change the resulting amino acid sequence of

involved in sensory perception, olfaction, and amino acid catabolism. They also showed that human genes that have been targeted by positive selection are significantly more likely to harbor variation associated with known genetic diseases. We here report the results of an analysis of 20,361 human and chimpanzee genes (of which 6,630 later were eliminated in a very conservative quality control), which includes the 7,645 genes analyzed by Clark et al. [10]. While the objective of the study by Clark et al. [10] was to find genes that have experienced accelerated evolution on the human lineage, using the mouse as an outgroup, the aim of the current study is to find genes that have been targeted by

Nielsen et al. (2005) *PLoS Bio* **3**:976-985.

Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup

Xiao-Jing Yu [a,b,c], Hong-Kun Zheng [d,e], Jun Wang [d,e,f], Wen Wang [a], Bing Su [a,b,*]

[a] *Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China*
[b] *Kunming Primate Research Center, Chinese Academy of Sciences, Kunming, Yunnan 650223, China*
[c] *Graduate School, Chinese Academy of Sciences, Beijing, China*
[d] *Beijing Genomics Institute, Chinese Academy of Sciences, Beijing, China*
[e] *The Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C, Denmark*
[f] *Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230, Odense M, Denmark*

Received 12 April 2006; accepted 23 May 2006
Available online 18 July 2006

Abstract

Comparative genetic analysis between human and chimpanzee may detect genetic divergences responsible for human-specific characteristics. Previous studies have identified a series of genes that potentially underwent Darwinian positive selection during human evolution. However, without a closely related species as outgroup, it is difficult to identify human-lineage-specific changes, which is critical in delineating the biological uniqueness of humans. In this study, we conducted phylogeny-based analyses of 2633 human brain-expressed genes using rhesus macaque as the outgroup. We identified 47 candidate genes showing strong evidence of positive selection in the human lineage. Genes with maximal expression in the brain showed a higher evolutionary rate in human than in chimpanzee. We observed that many immune-defense-related genes were under strong positive selection, and this trend was more prominent in chimpanzee than in human. We also demonstrated that rhesus macaque performed much better than mouse as an outgroup in identifying lineage-specific selection in humans.

Keywords: Positive selection; Adaptive evolution; Brain-expressed gene; Hominoids

Although nearly 99% of genomic sequences are identical, the biological divergence between human and chimpanzee is distinctive, especially in view of cognitive abilities [1,2]. Comparative genetic analysis of human and chimpanzee may detect genetic divergences responsible for the human-specific characteristics. Indeed, in the past several years, scientists have devoted great effort toward understanding the evolutionary

underwent Darwinian positive selection in humans [8]. They concluded that the putative positively selected genes in human were responsible mainly for several biological functions, including olfaction, sensory perception, and transportation. By analyzing 13,731 human–chimp orthologs, Nielsen et al. also showed that genes involved in sensory perception and immune system tend to evolve rapidly due to positive selection and

Yu et al. (2006) *Genomics* **88**:745-751.

No signal of adaptation in neural function from coding regions

Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios

Andrew G. Clark,[1] Stephen Glanowski,[3] Rasmus Nielsen,[2] Paul D. Thomas,[4] Anish Kejariwal,[4] Melissa A. Todd,[2] David M. Tanenbaum,[5] Daniel Civello,[6] Fu Lu,[5] Brian Murphy,[3] Steve Ferriera,[3] Gary Wang,[3] Xianqgun Zheng,[5] Thomas J. White,[6] John J. Sninsky,[6] Mark D. Adams,[5]* Michele Cargill[6]†

Even though human and chimpanzee gene sequences are nearly 99% identical, sequence comparisons can nevertheless be highly informative in identifying biologically important changes that have occurred since our ancestral lineages diverged. We analyzed alignments of 7645 chimpanzee gene sequences to their human and mouse orthologs. These three-species sequence alignments allowed us to identify genes undergoing natural selection along the human and chimp lineage by fitting models that include parameters specifying rates of synonymous and nonsynonymous nucleotide substitution. This evolutionary approach revealed an informative set of genes with significantly different patterns of substitution on the human lineage compared with the chimpanzee and mouse lineages. Partitions of genes into inferred biological classes identified accelerated evolution in several functional classes, including olfaction and nuclear transport. In addition to suggesting adaptive physiological differences between chimps and humans, human-accelerated genes are significantly more likely to underlie major known Mendelian disorders.

Although the human genome project will allow us to compare our genome to that of other primates and discover features that are uniquely human, there is no guarantee that such features are responsible for any of our unique biological attributes. To identify genes and biological processes that have been most altered by our recent evolutionary divergence from other primates, we need to fit the data to models of sequence divergence that allow us to distinguish between divergence caused by random drift and divergence driven by natural selection. Early observations of unexpectedly low levels of protein divergence between humans and chimpanzees led to the hypothesis that most of the evolutionary changes must have occurred at the level of gene regulation (*1*). Recently, much more extensive efforts at DNA sequencing in nonhuman primates has confirmed the very close evolutionary relationship between humans and chimps (*2*), with an average nucleotide divergence of just 1.2% (*3–5*). The role of protein divergence in causing morphological, physiological, and behavioral differences between these two species, however, remains unknown.

Here we apply evolutionary tests to identify genes and pathways from a new collection of more than 200,000 chimpanzee exonic sequences that show patterns of divergence consistent with natural selection along the human and chimpanzee lineages.

To construct the human-chimp-mouse alignments, we sequenced PCR amplifications using primers designed to essentially all human exons from one male chimpanzee, resulting in more than 20,000 human-chimp gene alignments spanning 18.5 Mb (*6–8*). To identify changes that are specific to the divergence in the human lineage, we compared the human-chimp aligned genes to their mouse ortholog. Inference of orthology involved a combination of reciprocal best matches and syntenic evidence between human and mouse gene annotations (*9*, *10*). This genome-wide set of orthologs underwent a series of filtering steps to remove ambiguities, orthologs with little sequence data, and genes with suspect annotation (*6*). The filtered ortholog set was compared to

[1]Molecular Biology and Genetics, [2]Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA. [3]Applied Biosystems, 45 West Gude Drive, Rockville, MD 20850, USA. [4]Protein Informatics, Celera Genomics, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [5]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [6]Celera Diagnostics, 1401 Harbor Bay Parkway, Alameda, CA 94502, USA.

*Present address: Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA.

†To whom correspondence should be addressed. E-mail: michele_cargill@celeradiagnostics.com

Clark et al. (2003) *Science* **302**:1960-1963.

A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees

Rasmus Nielsen[1,2*], Carlos Bustamante[1], Andrew G. Clark[3], Stephen Glanowski[4], Timothy B. Sackton[3], Melissa J. Hubisz[1], Adi Fledel-Alon[1], David M. Tanenbaum[5], Daniel Civello[6], Thomas J. White[6], John J. Sninsky[6], Mark D. Adams[5¤], Michele Cargill[6]

1 Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, 2 Center for Bioinformatics, University of Copenhagen, Denmark, 3 Molecular Biology and Genetics, Cornell University, Ithaca, New York, United States of America, 4 Applied Biosystems, Rockville, Maryland, United States of America, 5 Celera Genomics, Rockville, Maryland, United States of America, 6 Celera Diagnostics, Alameda, California, United States of America

Since the divergence of humans and chimpanzees about 5 million years ago, these species have undergone a remarkable evolution with drastic divergence in anatomy and cognitive abilities. At the molecular level, despite the small overall magnitude of DNA sequence divergence, we might expect such evolutionary changes to leave a noticeable signature throughout the genome. We here compare 13,731 annotated genes from humans to their chimpanzee orthologs to identify genes that show evidence of positive selection. Many of the genes that present a signature of positive selection tend to be involved in sensory perception or immune defenses. However, the group of genes that show the strongest evidence for positive selection also includes a surprising number of genes involved in tumor suppression and apoptosis, and of genes involved in spermatogenesis. We hypothesize that positive selection in some of these genes may be driven by genomic conflict due to apoptosis during spermatogenesis. Genes with maximal expression in the brain show little or no evidence for positive selection, while genes with maximal expression in the testis tend to be enriched with positively selected genes. Genes on the X chromosome also tend to show an elevated tendency for positive selection. We also present polymorphism data from 20 Caucasian Americans and 19 African Americans for the 50 annotated genes showing the strongest evidence for positive selection. The polymorphism analysis further supports the presence of positive selection in these genes by showing an excess of high-frequency derived nonsynonymous mutations.

Citation: Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3(6): e170.

Introduction

Genes, or regions of the genome, that have been affected by natural selection may show an excess of functionally important molecular changes, beyond what would be expected in the absence of selection. Genomic regions with such an excess of changes are said to have experienced positive selection, i.e., selection in favor of new genetic variants. The most common statistical technique for detecting positive selection takes advantage of the fact that mutations in coding regions of genes come in two classes: nonsynonymous mutations that change the resulting amino acid sequence of

involved in sensory perception, olfaction, and amino acid catabolism. They also showed that human genes that have been targeted by positive selection are significantly more likely to harbor variation associated with known genetic diseases. We here report the results of an analysis of 20,361 human and chimpanzee genes (of which 6,630 later were eliminated in a very conservative quality control), which includes the 7,645 genes analyzed by Clark et al. [10]. While the objective of the study by Clark et al. [10] was to find genes that have experienced accelerated evolution on the human lineage, using the mouse as an outgroup, the aim of the current study is to find genes that have been targeted by

Nielsen et al. (2005) *PLoS Bio* **3**:976-985.

Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup

Xiao-Jing Yu [a,b,c], Hong-Kun Zheng [d,e], Jun Wang [d,e,f], Wen Wang [a], Bing Su [a,b,*]

[a] *Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China*
[b] *Kunming Primate Research Center, Chinese Academy of Sciences, Kunming, Yunnan 650223, China*
[c] *Graduate School, Chinese Academy of Sciences, Beijing, China*
[d] *Beijing Genomics Institute, Chinese Academy of Sciences, Beijing, China*
[e] *The Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C, Denmark*
[f] *Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230, Odense M, Denmark*

Received 12 April 2006; accepted 23 May 2006
Available online 18 July 2006

Abstract

Comparative genetic analysis between human and chimpanzee may detect genetic divergences responsible for human-specific characteristics. Previous studies have identified a series of genes that potentially underwent Darwinian positive selection during human evolution. However, without a closely related species as outgroup, it is difficult to identify human-lineage-specific changes, which is critical in delineating the biological uniqueness of humans. In this study, we conducted phylogeny-based analyses of 2633 human brain-expressed genes using rhesus macaque as the outgroup. We identified 47 candidate genes showing strong evidence of positive selection in the human lineage. Genes with maximal expression in the brain showed a higher evolutionary rate in human than in chimpanzee. We observed that many immune-defense-related genes were under strong positive selection, and this trend was more prominent in chimpanzee than in human. We also demonstrated that rhesus macaque performed much better than mouse as an outgroup in identifying lineage-specific selection in humans.

Keywords: Positive selection; Adaptive evolution; Brain-expressed gene; Hominoids

Although nearly 99% of genomic sequences are identical, the biological divergence between human and chimpanzee is distinctive, especially in view of cognitive abilities [1,2]. Comparative genetic analysis of human and chimpanzee may detect genetic divergences responsible for the human-specific characteristics. Indeed, in the past several years, scientists have devoted great effort toward understanding the evolutionary

underwent Darwinian positive selection in humans [8]. They concluded that the putative positively selected genes in human were responsible mainly for several biological functions, including olfaction, sensory perception, and transportation. By analyzing 13,731 human–chimp orthologs, Nielsen et al. also showed that genes involved in sensory perception and immune system tend to evolve rapidly due to positive selection and

Yu et al. (2006) *Genomics* **88**:745-751.

No enrichment for brain-expressed genes

No enrichment for neural developmental or neural function

Over-represented functional categories: 5’ flanking region

category	# genes	*p*-value on: human branch	chimp branch
protein folding	70	0.0067	0.77
other neuronal activity	31	0.013	0.039
neurogenesis	133	0.013	0.032
glycolysis	21	0.014	0.72
carbohydrate metabolism	210	0.020	0.017
ectoderm development	169	0.020	0.11
mesoderm development	161	0.024	0.17
synaptic transmission	25	0.025	0.34
vision	64	0.025	0.15
oncogene	25	0.045	0.46
anion transport	31	0.049	0.17

Human branch, *alt2-null2* model contrast. Biological Process categories based on Panther Classification System (www.pantherdb.org). *p*-value = one-tailed Mann-Whitney test for human or chimpanzee branch. *p*-value < 0.01, *p*-value < 0.03, *p*-value < 0.03

Over-represented functional categories: 5’ flanking region

	category	# genes	*p*-value on: human branch	*p*-value on: chimp branch
	protein folding	70	0.0067	0.77
●	other neuronal activity	31	0.013	0.039
●	neurogenesis	133	0.013	0.032
	glycolysis	21	0.014	0.72
	carbohydrate metabolism	210	0.020	0.017
●	ectoderm development	169	0.020	0.11
	mesoderm development	161	0.024	0.17
●	synaptic transmission	25	0.025	0.34
	vision	64	0.025	0.15
	oncogene	25	0.045	0.46
	anion transport	31	0.049	0.17

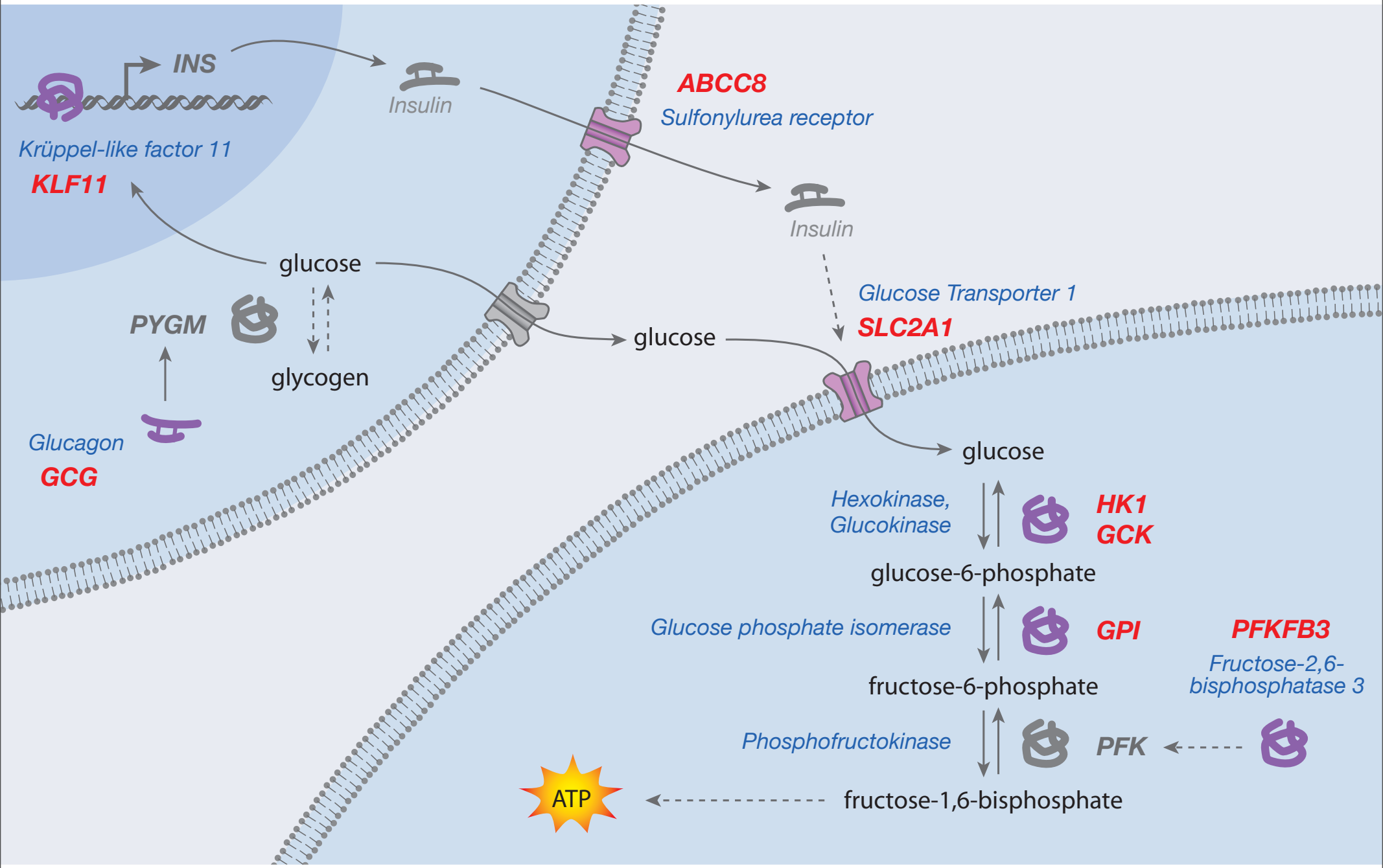
Human branch, *alt2-null2* model contrast. Biological Process categories based on Panther Classification System (www.pantherdb.org). *p*-value = one-tailed Mann-Whitney test for human or chimpanzee branch. *p*-value < 0.01, *p*-value < 0.03, *p*-value < 0.03

Over-represented functional categories: 5’ flanking region

	category	# genes	*p*-value on: human branch	*p*-value on: chimp branch
	protein folding	70	0.0067	0.77
●	other neuronal activity	31	0.013	0.039
●	neurogenesis	133	0.013	0.032
●	glycolysis	21	0.014	0.72
●	carbohydrate metabolism	210	0.020	0.017
●	ectoderm development	169	0.020	0.11
	mesoderm development	161	0.024	0.17
●	synaptic transmission	25	0.025	0.34
	vision	64	0.025	0.15
	oncogene	25	0.045	0.46
	anion transport	31	0.049	0.17

Human branch, *alt2-null2* model contrast. Biological Process categories based on Panther Classification System (www.pantherdb.org). *p*-value = one-tailed Mann-Whitney test for human or chimpanzee branch. *p*-value < 0.01, *p*-value < 0.03, *p*-value < 0.03

Positive selection on a functional cluster: sugar metabolism



Why positive selection on regulation of sugar metabolism?

Shift in diet

- Humans are an outlier among the great apes in terms of diet
- Other great apes consume primarily fruits, nuts, and roots
- Humans rely on a more diverse source of foods, especially more meat

Metabolic demands of a larger brain

- Human brain volume much larger that of human-chimp ancestor
- Metabolic demands scale linearly with volume
- Modifications in metabolism necessary to compensate for larger brains

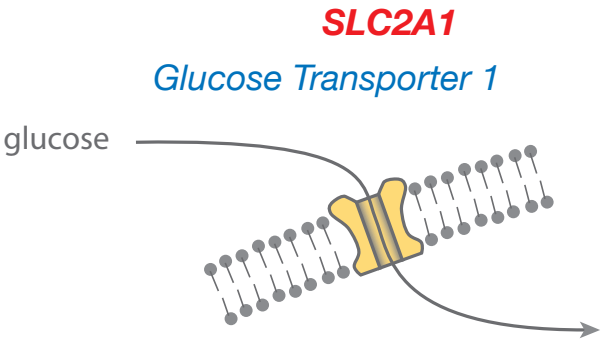
Why positive selection on regulation of sugar metabolism?

Shift in diet

- Humans are an outlier among the great apes in terms of diet
- Other great apes consume primarily fruits, nuts, and roots
- Humans rely on a more diverse source of foods, especially more meat

Metabolic demands of a larger brain

- Human brain volume much larger that of human-chimp ancestor
- Metabolic demands scale linearly with volume
- Modifications in metabolism necessary to compensate for larger brains

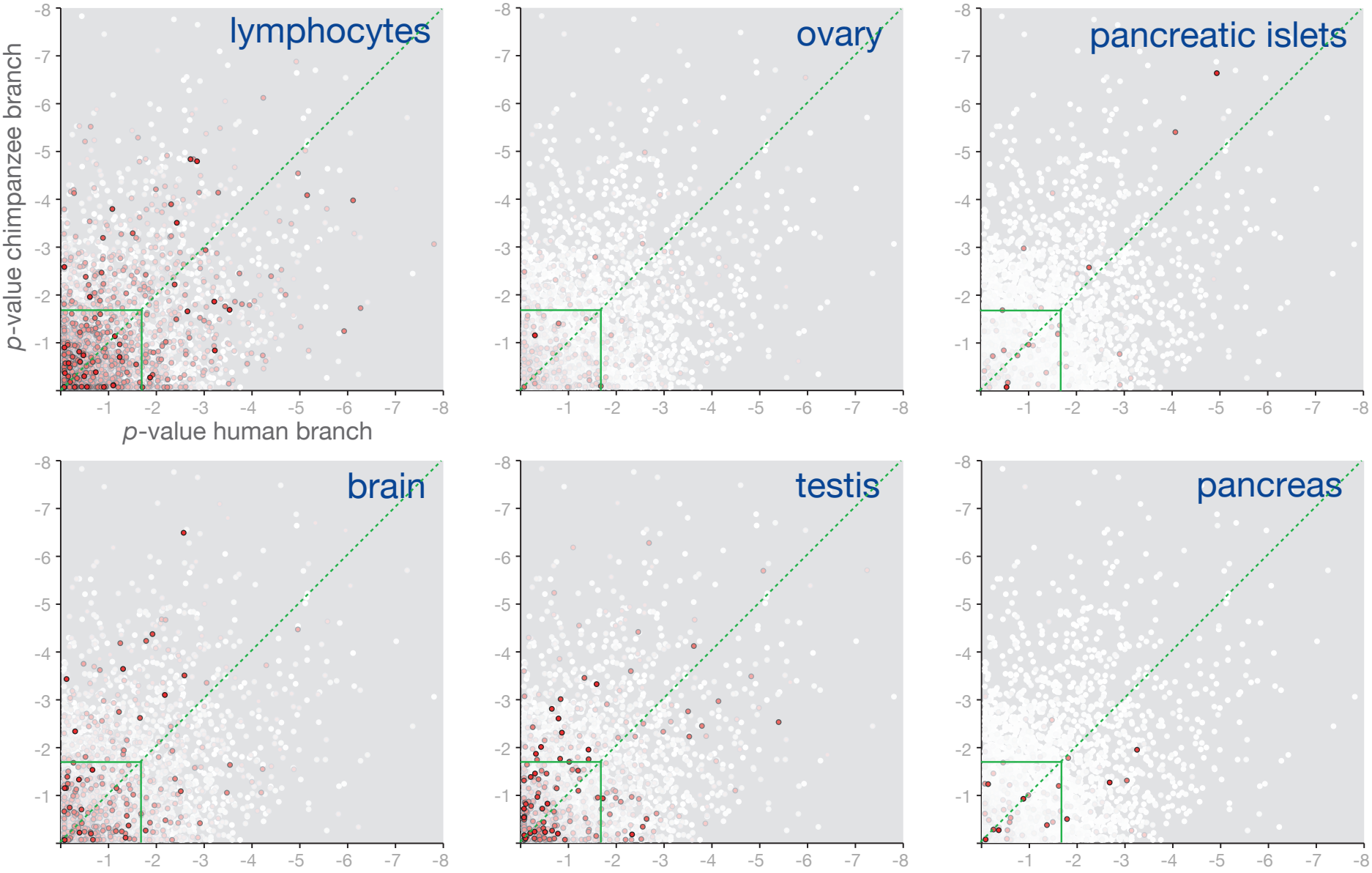


Major glucose transporter across blood-brain barrier

Association: noninsulin-dependent diabetes mellitus

Hemizygous: acquired microcephaly, cognitive impairments

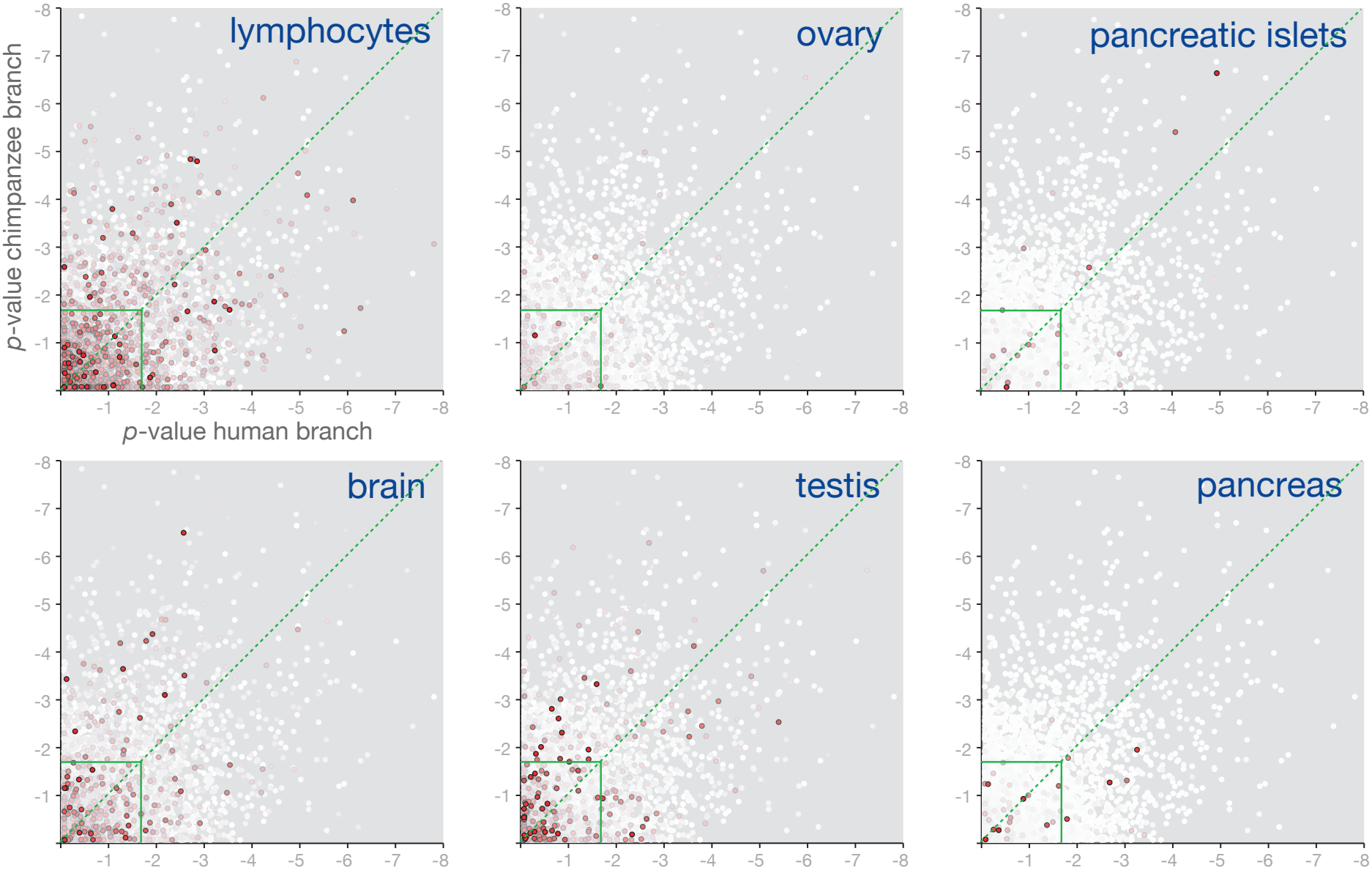
Branch-specific positive selection and expression specificity



Expression of 6,316 genes in 73 tissues from Novartis Human Gene Expression Atlas.

Expression specificity: *unique* *ubiquitous*

Branch-specific positive selection and expression specificity



Expression of 6,316 genes in 73 tissues from Novartis Human Gene Expression Atlas.

Expression specificity: *unique* *ubiquitous*

Why selection on the chimpanzee branch?

Chimpanzee art, painted by Congo in 1957

King and Wilson's proposal

Phenotypic differences between humans and chimpanzees are due primarily to changes in the regulation of gene expression

King, M-C and A Wilson (1975) *Science* 188:107

More positive selection on 5’ flanking than coding regions

partition	*source*	*proportion*	*counts*
5’ flanking	our analysis	0.0041	26/6316
coding	Clark et al.	0.0000	0/7645
5’ flanking	our analysis	0.0140	88/6316
coding	Nielsen et al.	0.0019	21/11159

Counts based on q-values < 0.05

Sources: Clark et al. (2003) *Science* 302:1960; Nielsen et al. (2005) *PLoS Biology* 3:976.

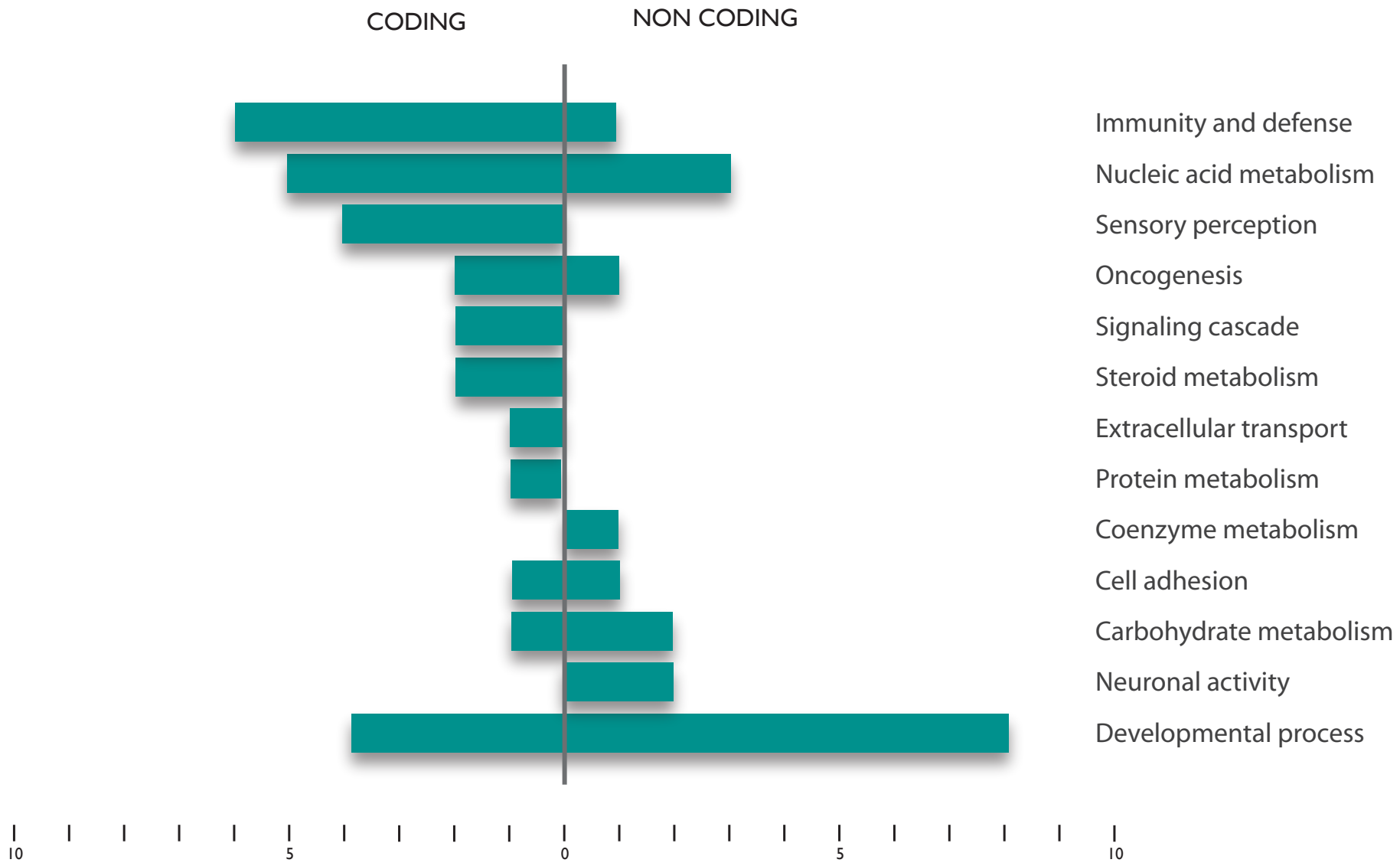
More positive selection on 5’ flanking than coding regions

partition	*source*	*proportion*	*counts*
5’ flanking	our analysis	0.0041	26/6316
coding	Clark et al.	0.0000	0/7645
5’ flanking	our analysis	0.0140	88/6316
coding	Nielsen et al.	0.0019	21/11159

Counts based on q-values < 0.05

Sources: Clark et al. (2003) *Science* 302:1960; Nielsen et al. (2005) *PLoS Biology* 3:976.

Qualitative differences in selection on coding/regulatory regions



Meta-analysis of published genome-wide scans for selection: three coding region and two non-coding region.
All Panther Biological Process categories with meta p-value < 0.05, sum of log-weighted meta p-values

Part 1 summary

- Lots of positive selection on regulatory sequences
- Neural, metabolic functions prominently affected
- Many new candidate genes for further study

screening for more candidate genes

testing for functional differences

connecting the dots: 3 case studies

Studying *cis*-regulatory evolution

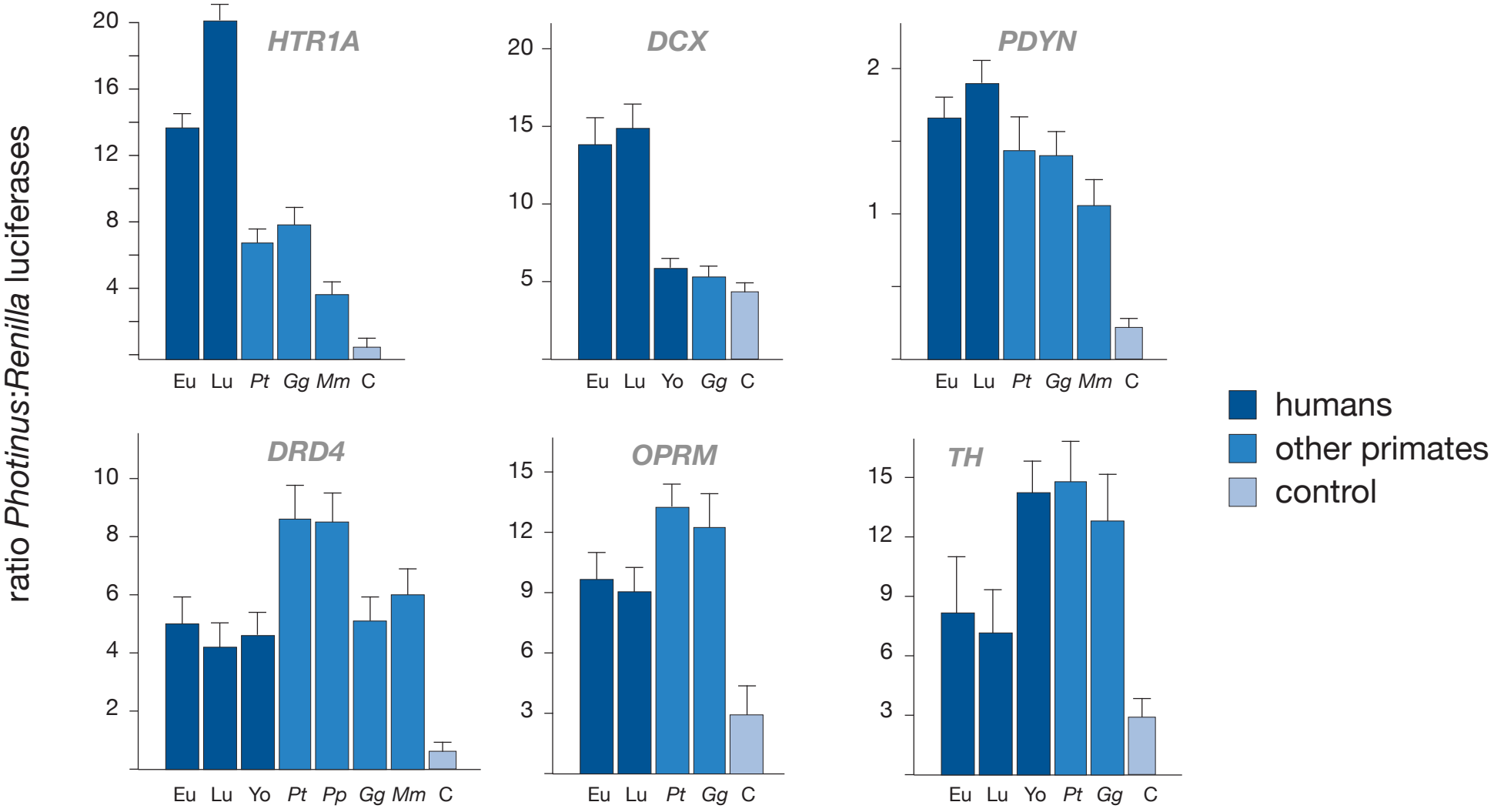
the bad news: *cis*-regulatory sequences are cryptic and spatially unbounded

the good news: functional tests are easy and uniformly applicable

Testing for differences in gene expression

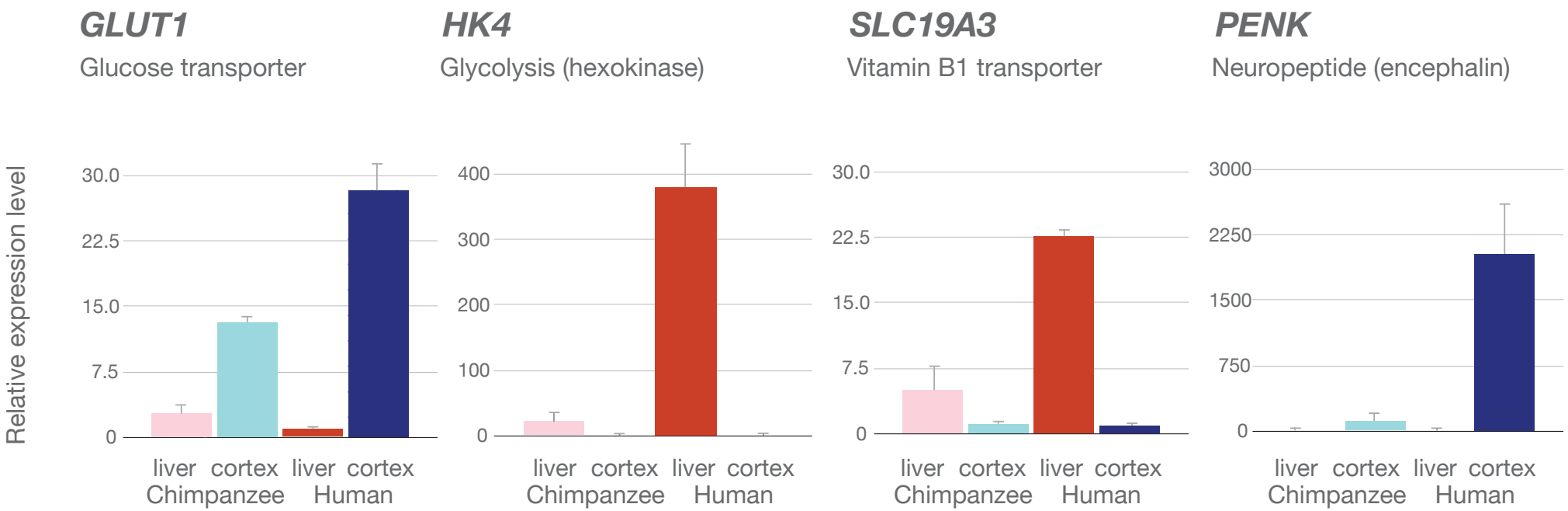
method	context	scale	genetics?	sensitivity	precision
reporter gene assay	*in vitro*	1-10s	yes	high	good
quantitative PCR	*in vivo*	1-100s	no/yes	high	good
pyrosequencing	*in vivo*	1-100s	yes	high	good
microarrays	*in vivo*	10,000s	no	low	fair
SAGE profiling	*in vivo*	10,000s	no	high	good

Expression differences by reporter gene assays



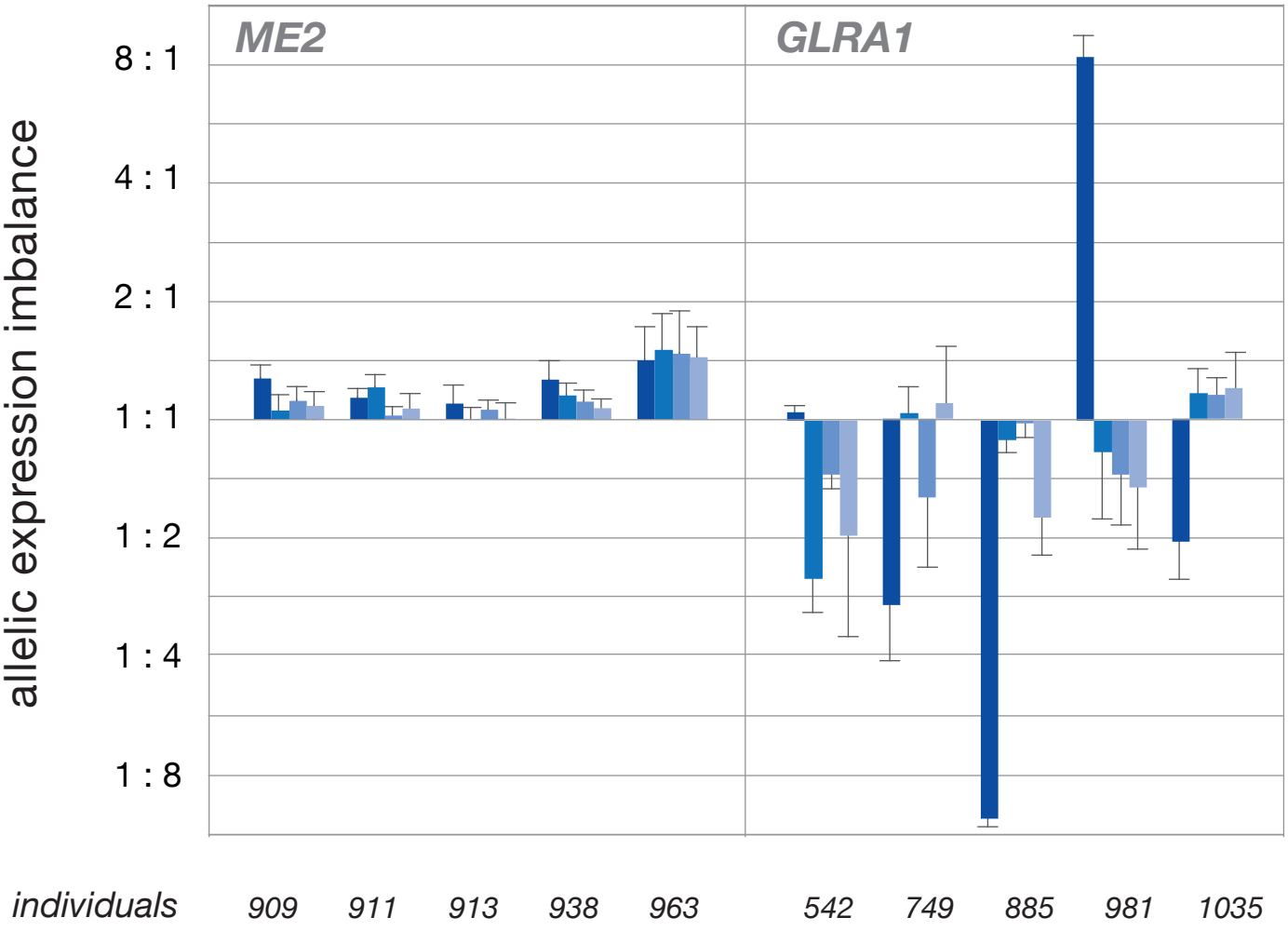
Six brain-expressed genes whose 5' flanking regions show strong signals of positive selection. Dual Luciferase Assay in IMR-32 or SH-SY5Y cells; bars are means of 8 technical replicates.

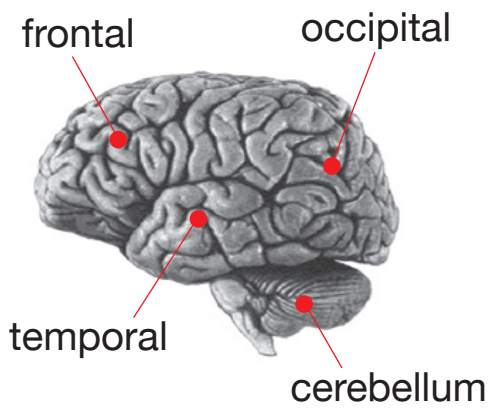
Expression differences between species by qPCR

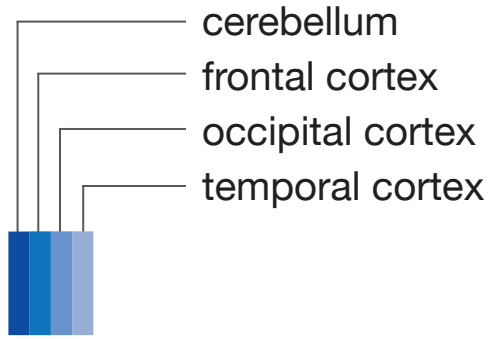


Four genes of interest: 3 with strong signatures of positive selection, 4th is a candidate gene.
Quantitative PCR; values are means of 4 individuals and 6 technical replicates.
Error bars are standard deviations on all 24 measurements for each tissue x species.

Expression differences between alleles in the same individual







Two brain-expressed genes whose 5' flanking regions show strong signals of positive selection. Allelic imbalance measured by pyrosequencing; bars are means of 8 technical replicates.

Part 2 summary

- Same functional tests can be applied to almost any gene
- Tests for functional changes in regulation are informative
- Positively-selected genes often show functional changes

screening for more candidate genes

testing for functional differences

connecting the dots: 3 case studies

LCT

Encodes lactase-phlorizin hydrolase

Dual-function enzyme

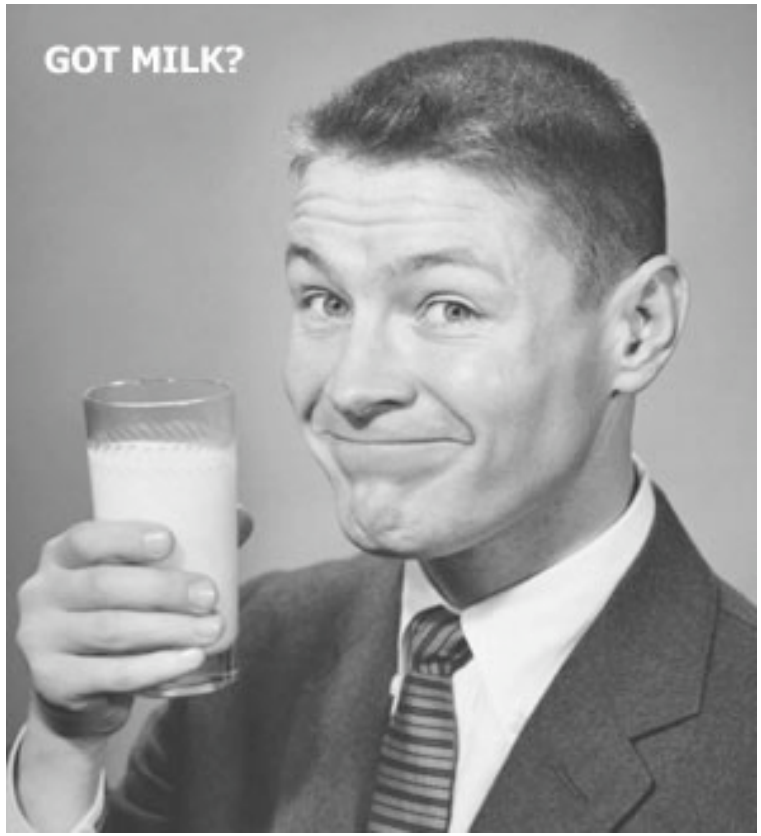
Digests lactose

Responsible for lactase persistence

Ability to metabolize lactose as adults

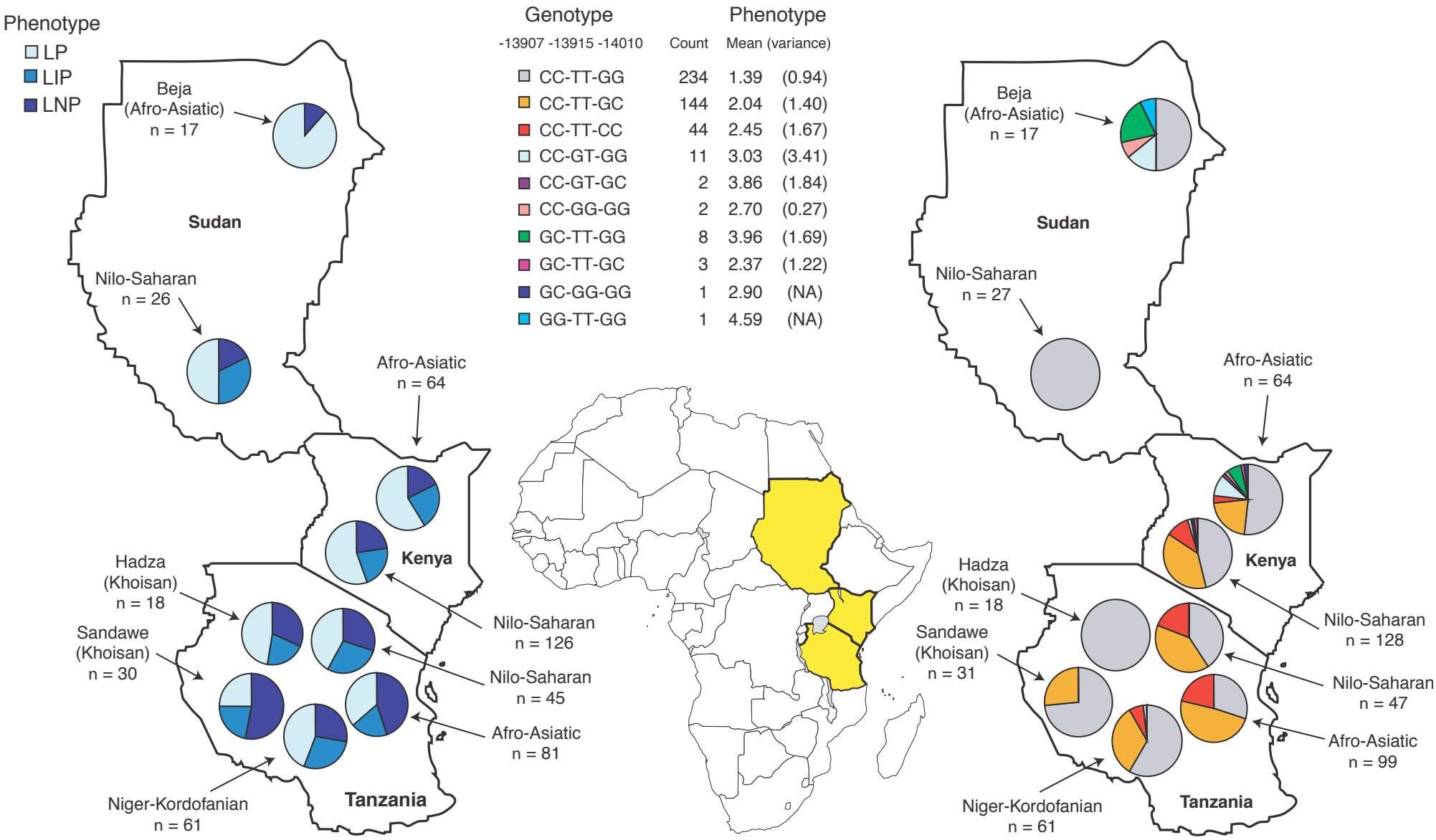
Evolved very recently (< 20,000 years)

Associated with pastoralism



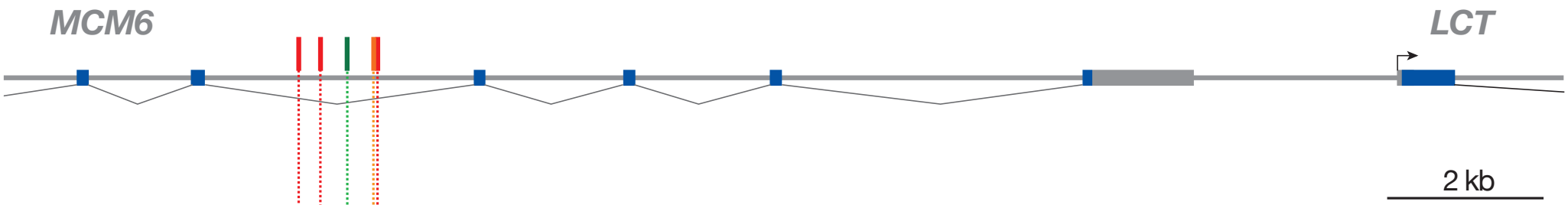
An enzyme associated with a uniquely human metabolic function

Lactase persistence: genotype and phenotype



Collaboration with: Sarah Tishkoff, Jonathan Pritchard, Panos Deloukas

Genetic basis: implicated SNPs



-13910 C/T "European SNP" (Enattah et al. 2002)

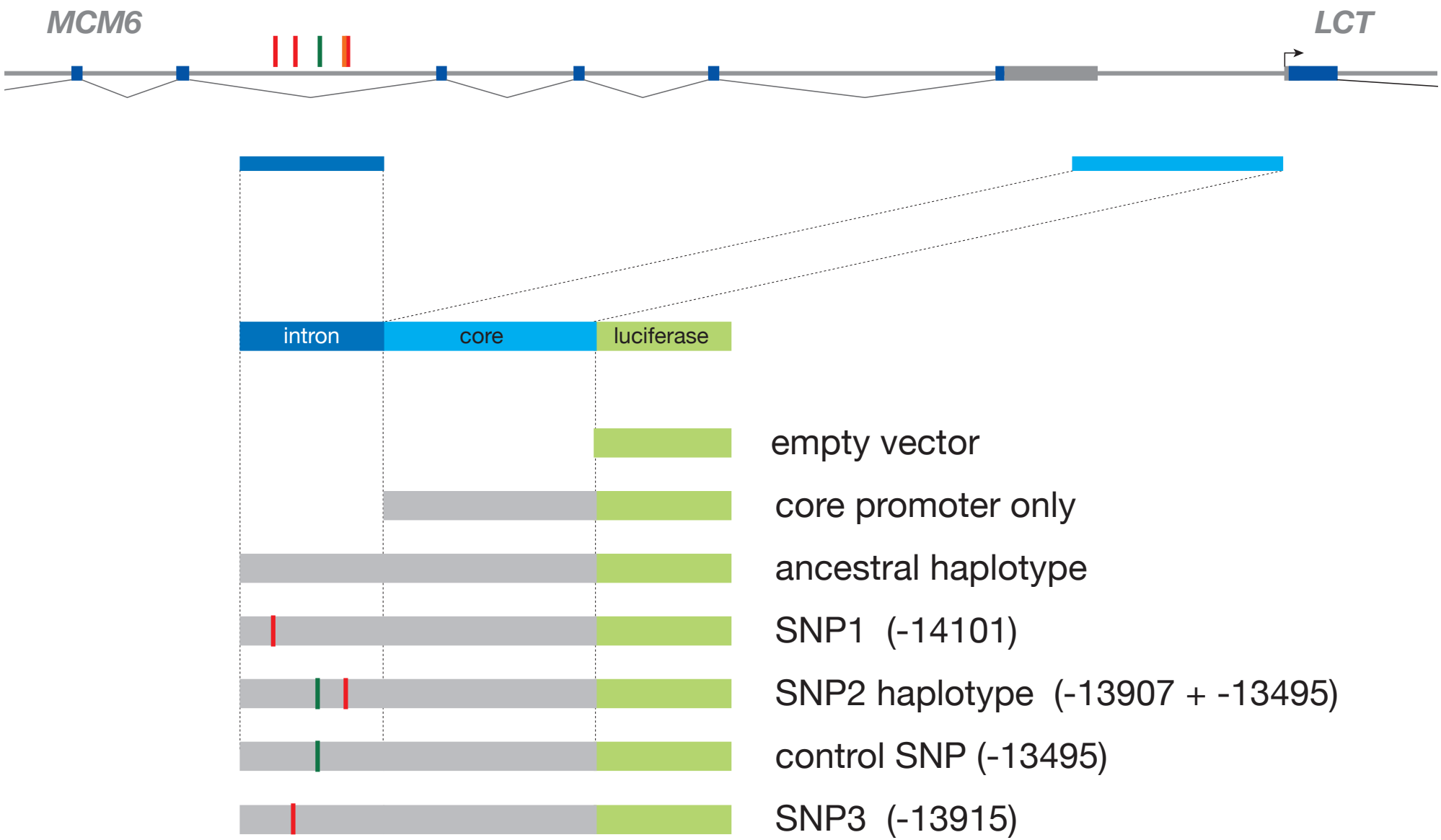
-14101 G/C SNP1: Kenyan, Tanzanian (Rendile, Sandawe)

-13907 C/G SNP2: Northern Sudanese (Beja)

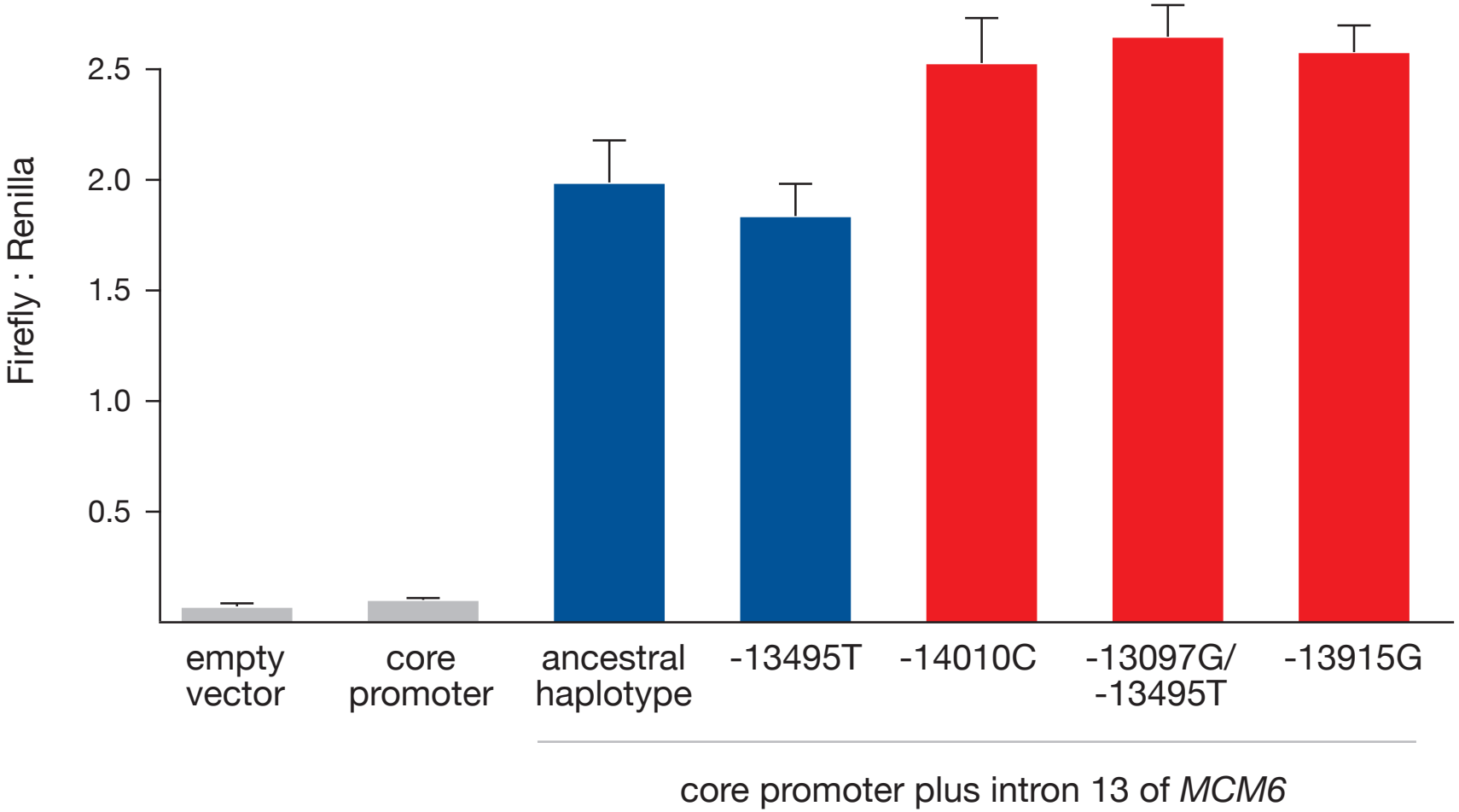
-13915 T/G SNP3: Ethiopian (Gabre)

-13495 C/T control SNP: no association with LP

Constructs for testing implicated SNPs

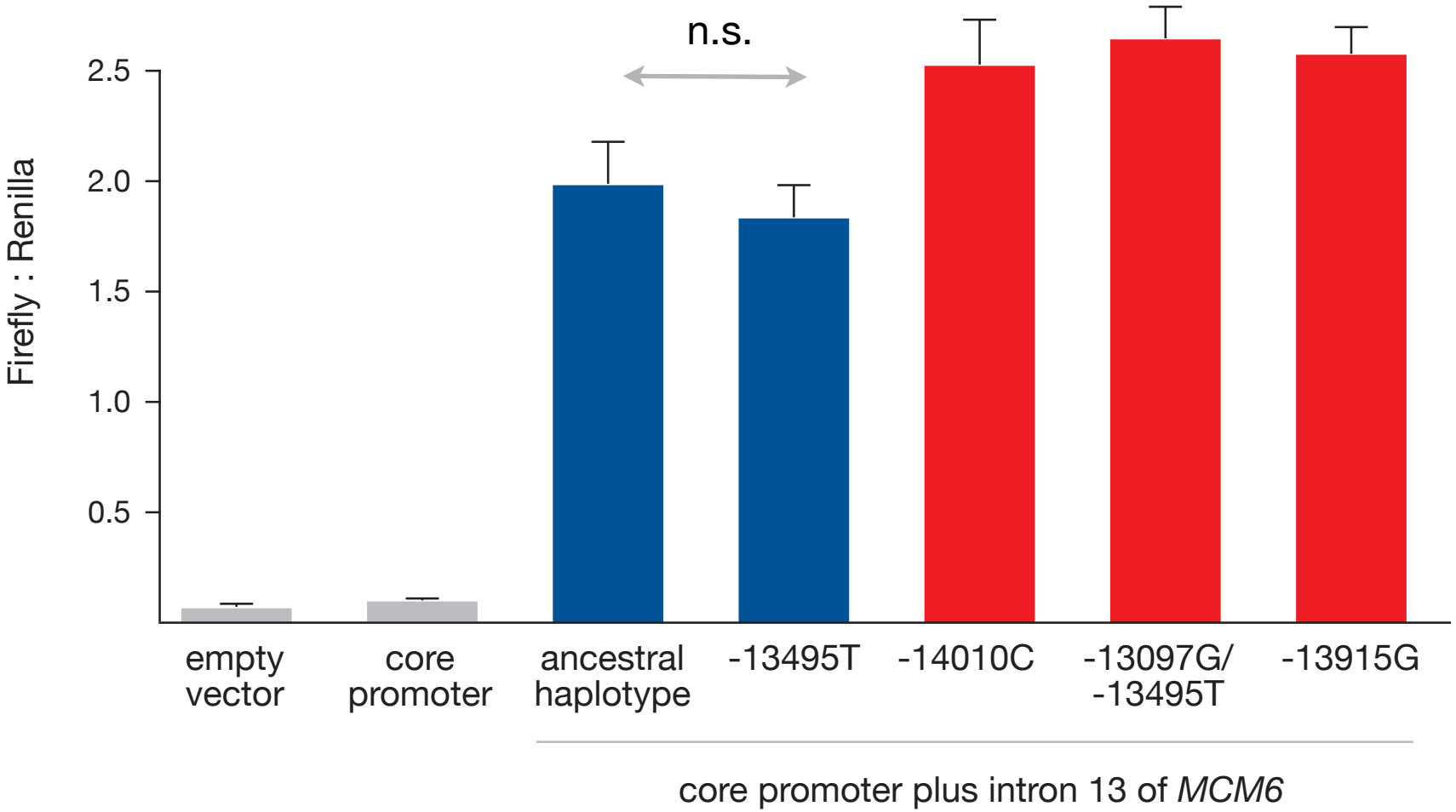


All three implicated SNPs elevate transcription



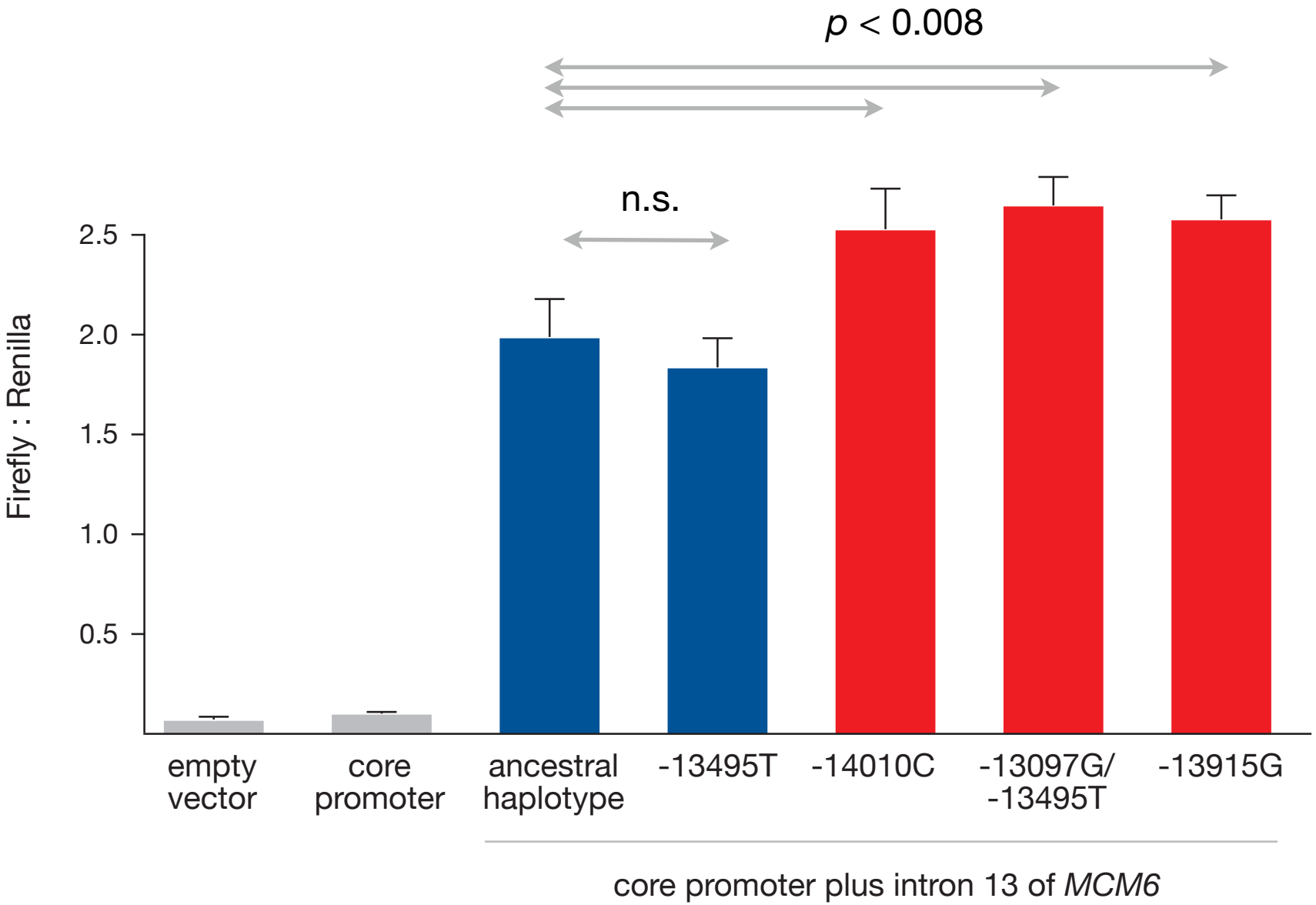
Dual-Luciferase assay. Bars are means of 12 technical replicates.

All three implicated SNPs elevate transcription



Dual-Luciferase assay. Bars are means of 12 technical replicates.

All three implicated SNPs elevate transcription



Dual-Luciferase assay. Bars are means of 12 technical replicates.

The ecological agent

Hans-Peter Harmsen

DARC

Encodes chemokine receptor Duffy

Gene formerly known as *FY*

Signaling within the immune system

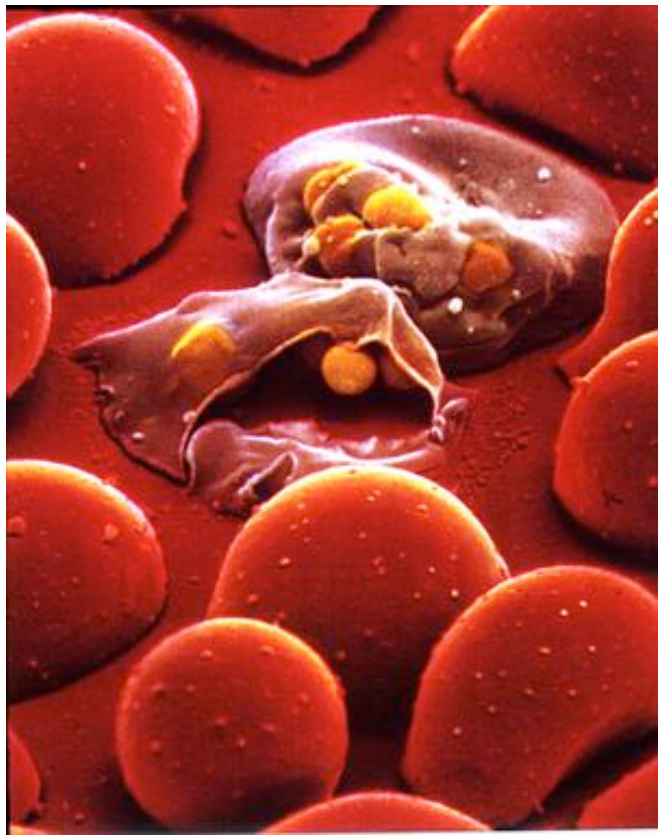
Resistance to *Plasmodium vivax* infection

SNP -32 abolishes erythrocyte expression

Monogenic trait, semi-dominant

Allele fixed in regions of malarial endemism

Clear signature of natural selection



University of Florida School of Medicine

A dramatic example of adaptation in cis-regulatory function

Savannah baboons in the Amboseli basin, Kenya

Project of Stuart Altmann, Jean Altmann, and Susan Alberts

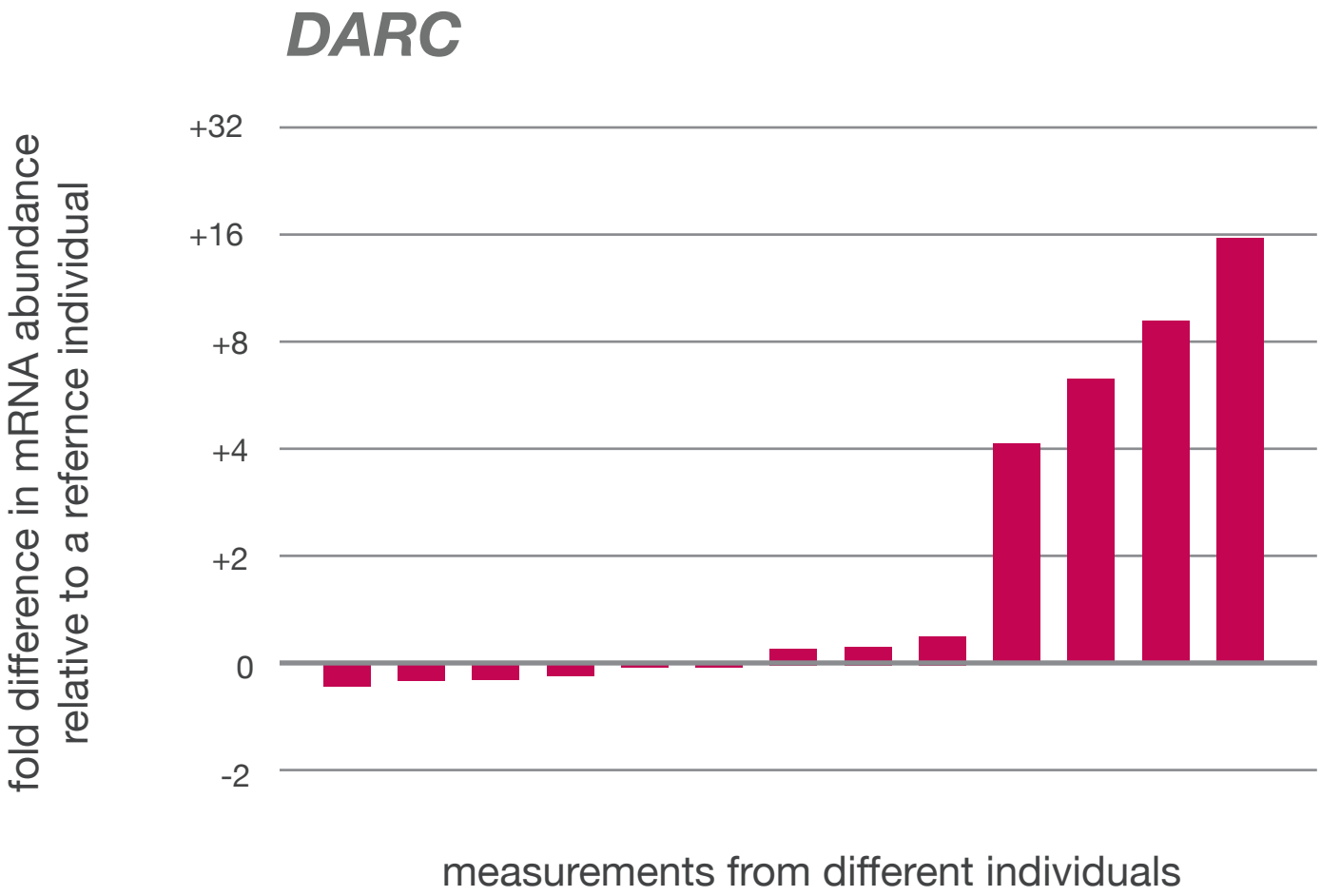
Continuous field observation and detailed data from 1971-present

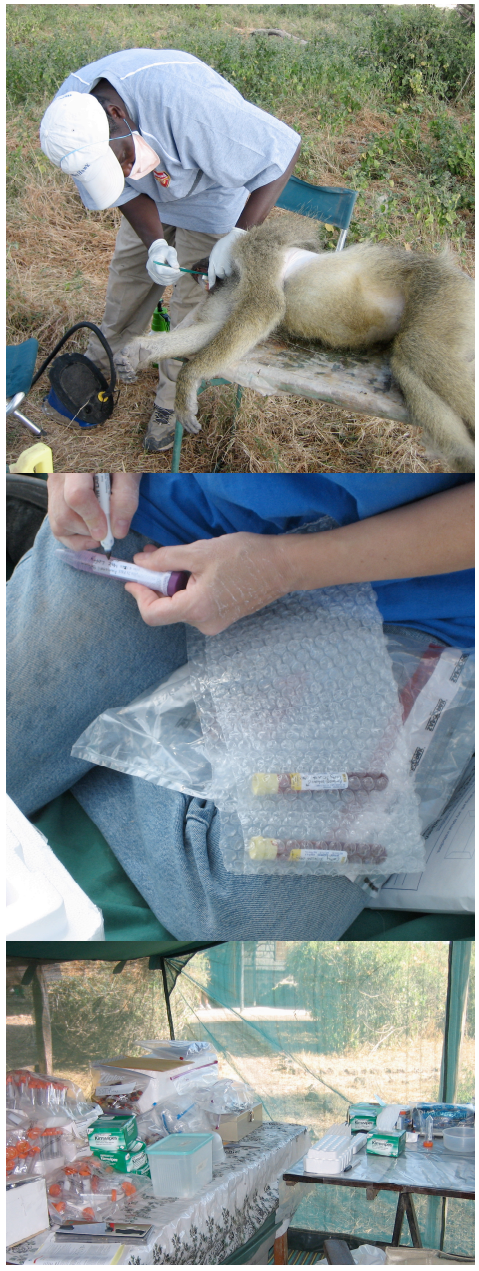
Data from >1000 individuals, many known pedigrees (3-5 generations)

Genomic DNA from >600 individuals

cDNA from peripheral blood and skin biopsies from >80 individuals

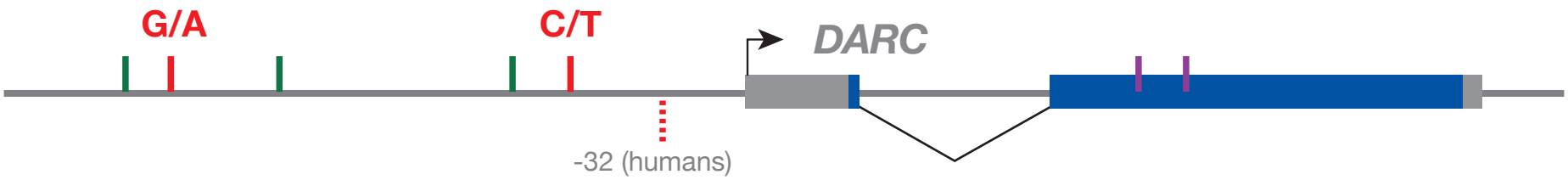
In vivo differences in transcript abundance in wild individuals

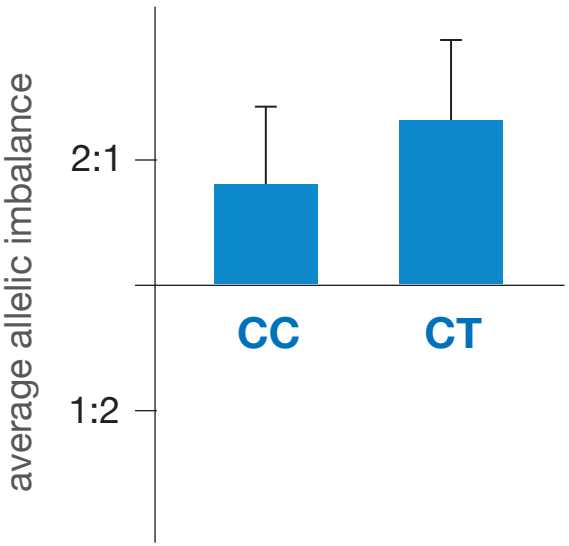




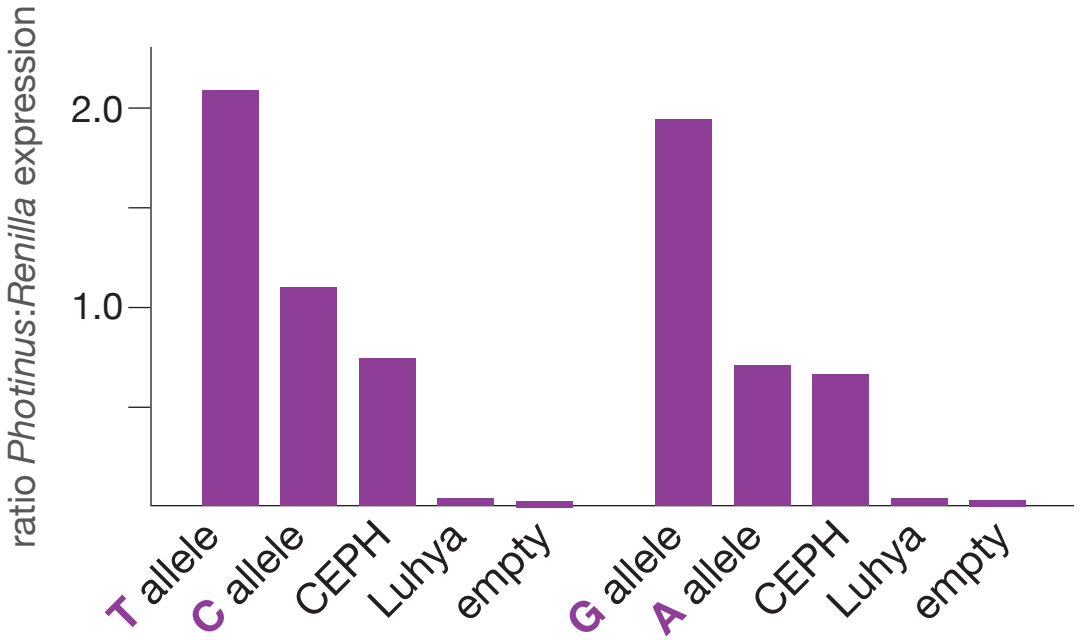
Jenny Tung

Two 5’ SNPs near *DARC* influence expression levels



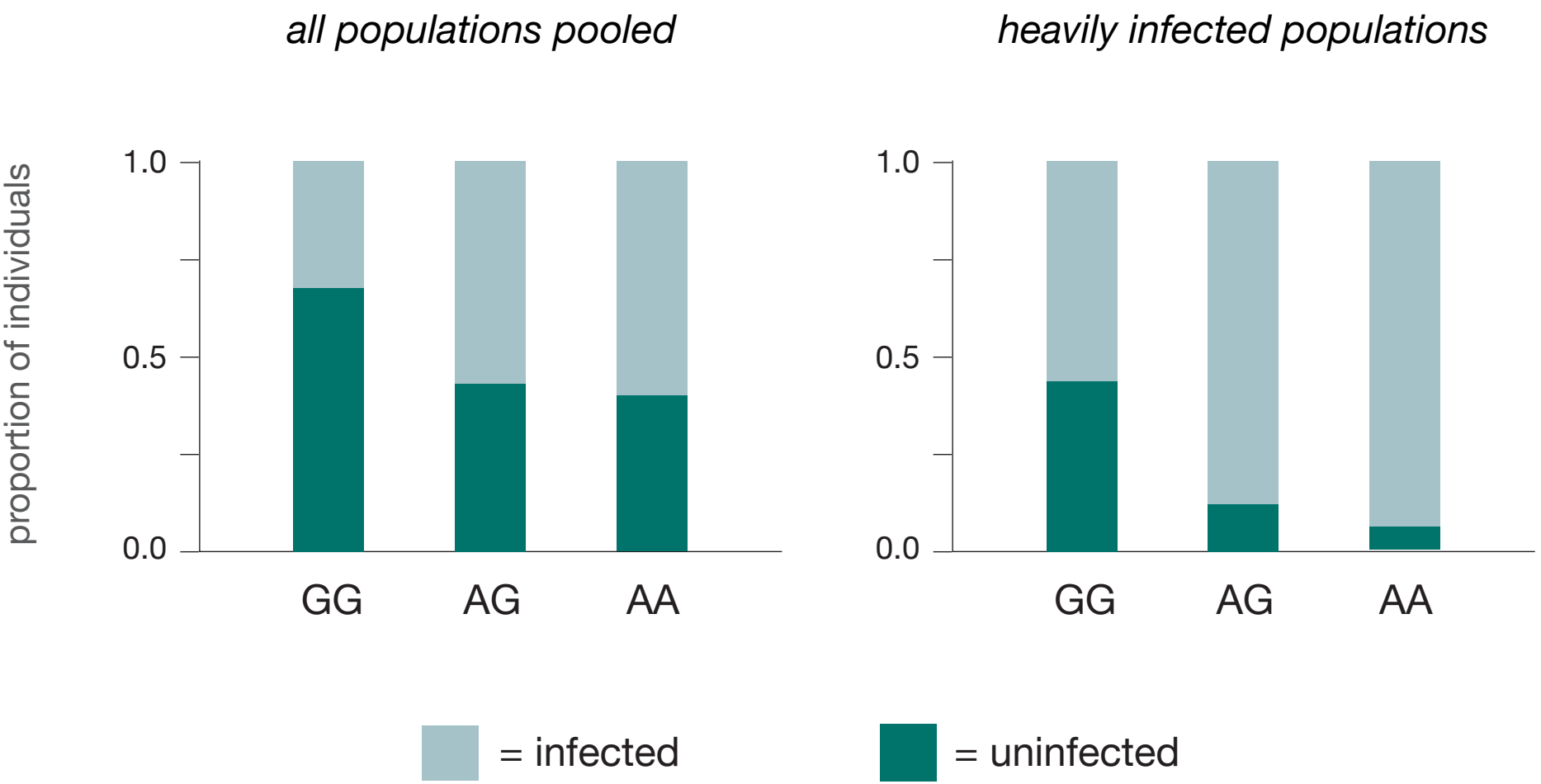


Genotype of proximal SNP predicts *in vivo* allelic imbalance ($p < 0.047$)



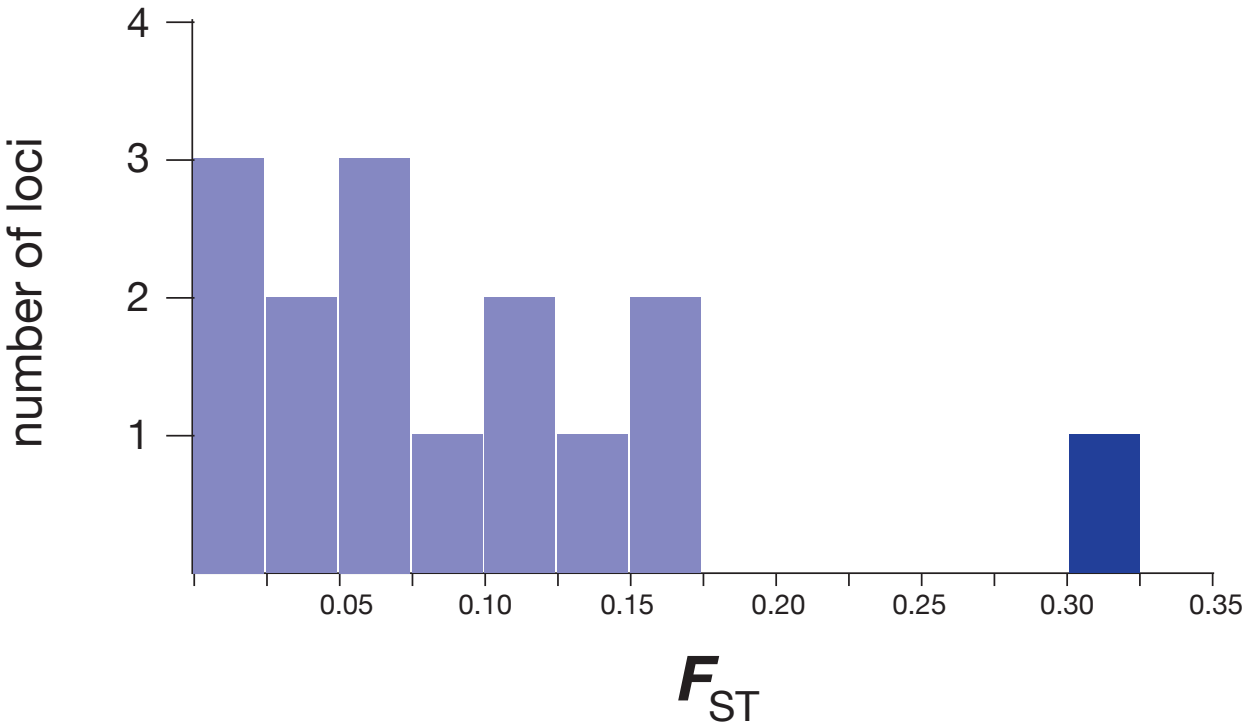
Relative expression levels, fuctional assays *in vitro* in HEL cells (all between construct differences significant)

Distal SNP genotype correlated with *Hepatocystis* infection



Natural infections of baboons by: *Plasmodium knowlesi* and closely related haematoprotozoan parasites within the in genera *Babesia*, *Entopolyploides*, and *Hepatocystis*.

Signatures of selection on the 5’ flanking region of *DARC*



Significant differentiation ($p < 0.001$) among three wild baboon populations: Amboseli (Kenya), Masai Mara (Kenya), and Mikumi (Tanzania), suggesting local directional selection.

Tajima’s D also significantly elevated (D = 1.755, $p < 0.024$), suggesting balancing selection.

DRD4

Encodes a dopamine receptor

Metabotropic G protein-coupled receptor

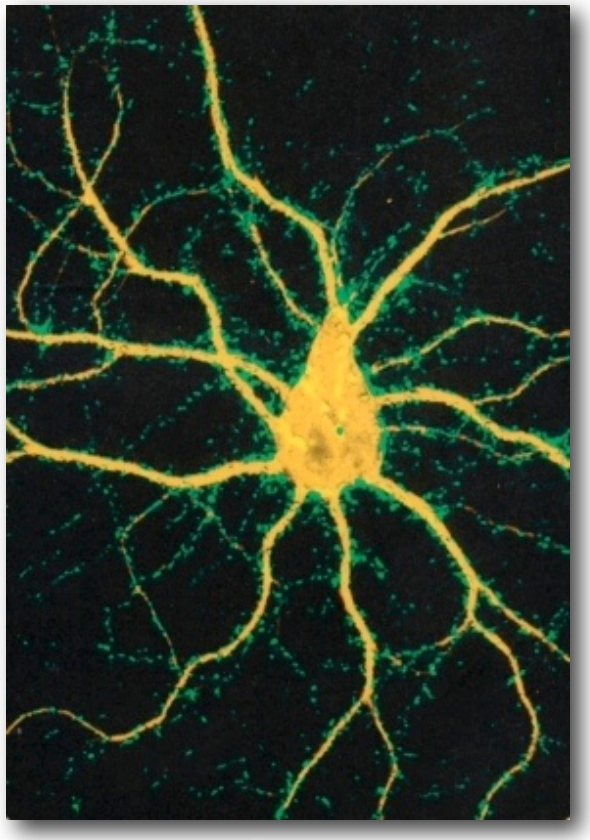
Inhibitory to post-synaptic neuron

Genetically associated with personality traits

Novelty-seeking, impulsiveness, risk-taking

Attention-deficit hyperactivity disorder

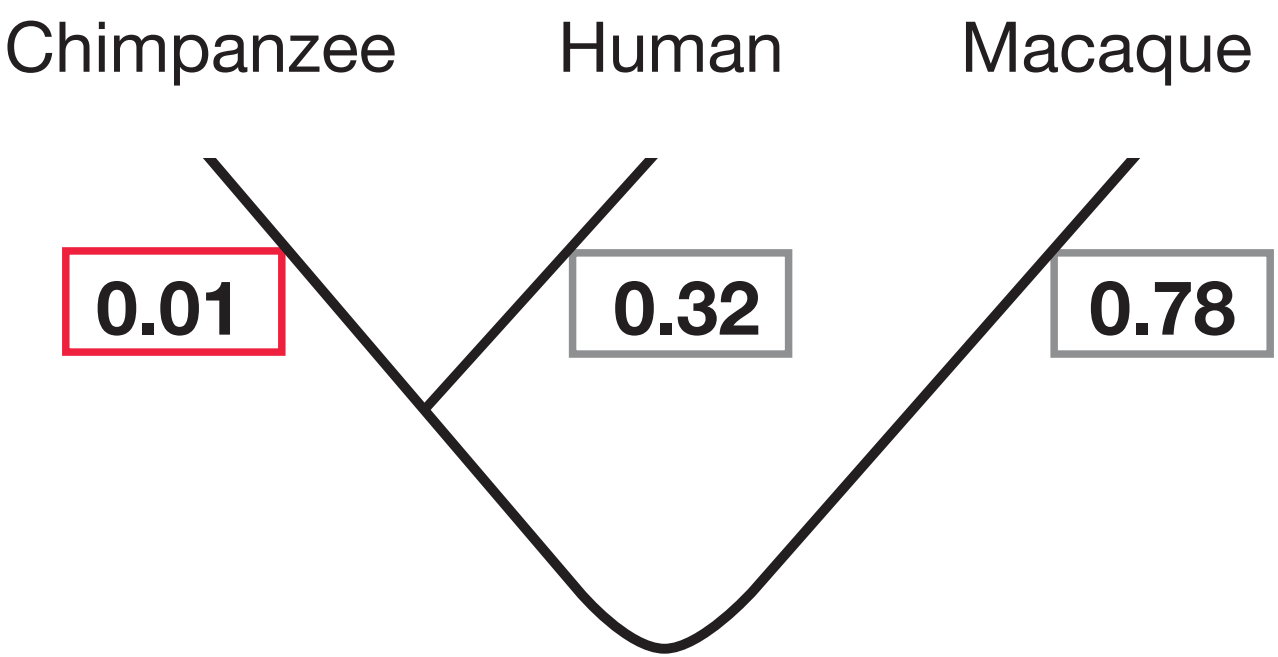
Schizophrenia



Peter Vanhoenacker

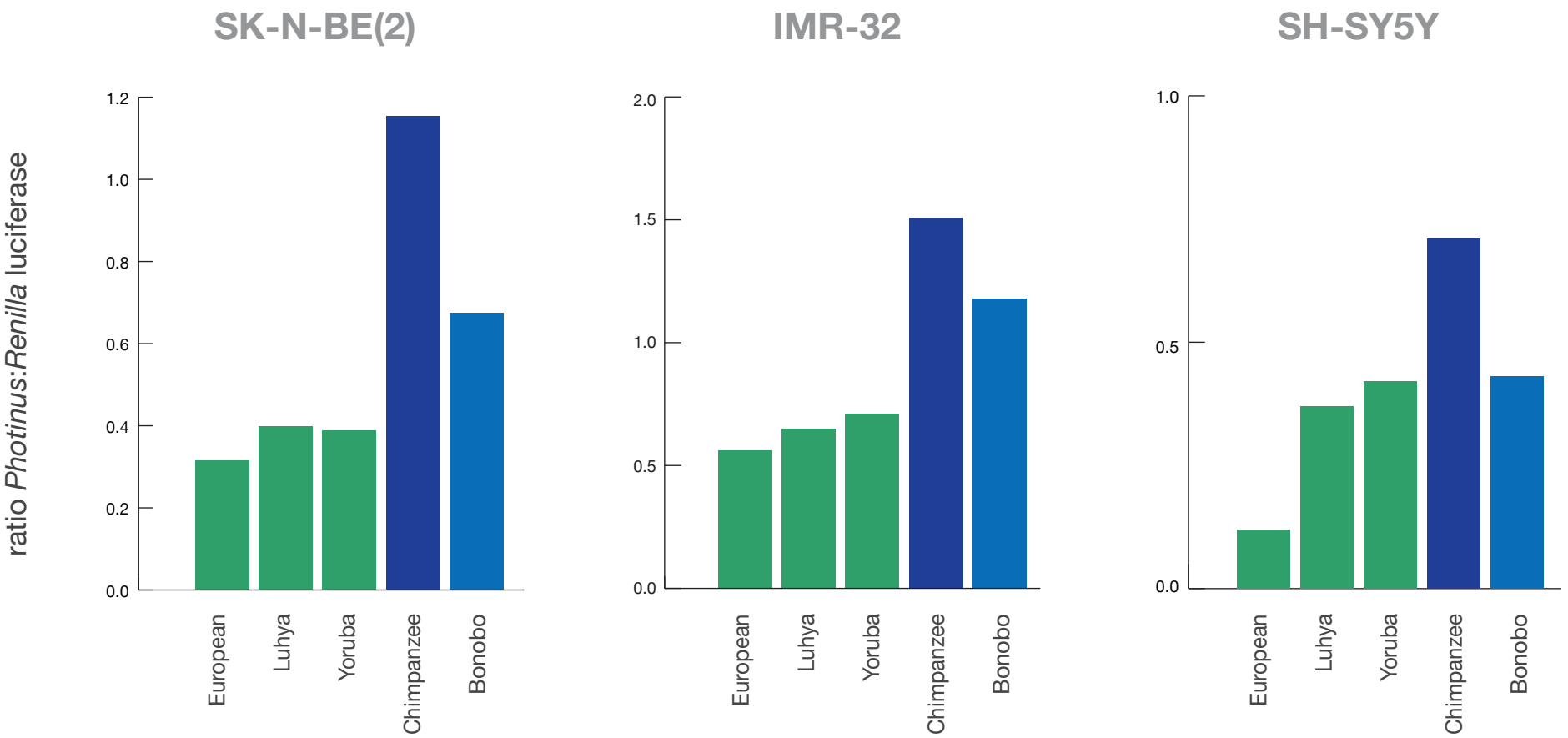
A protein associated with variation in cognition in humans

Branch-specific positive selection on *DRD4* 5' flanking region



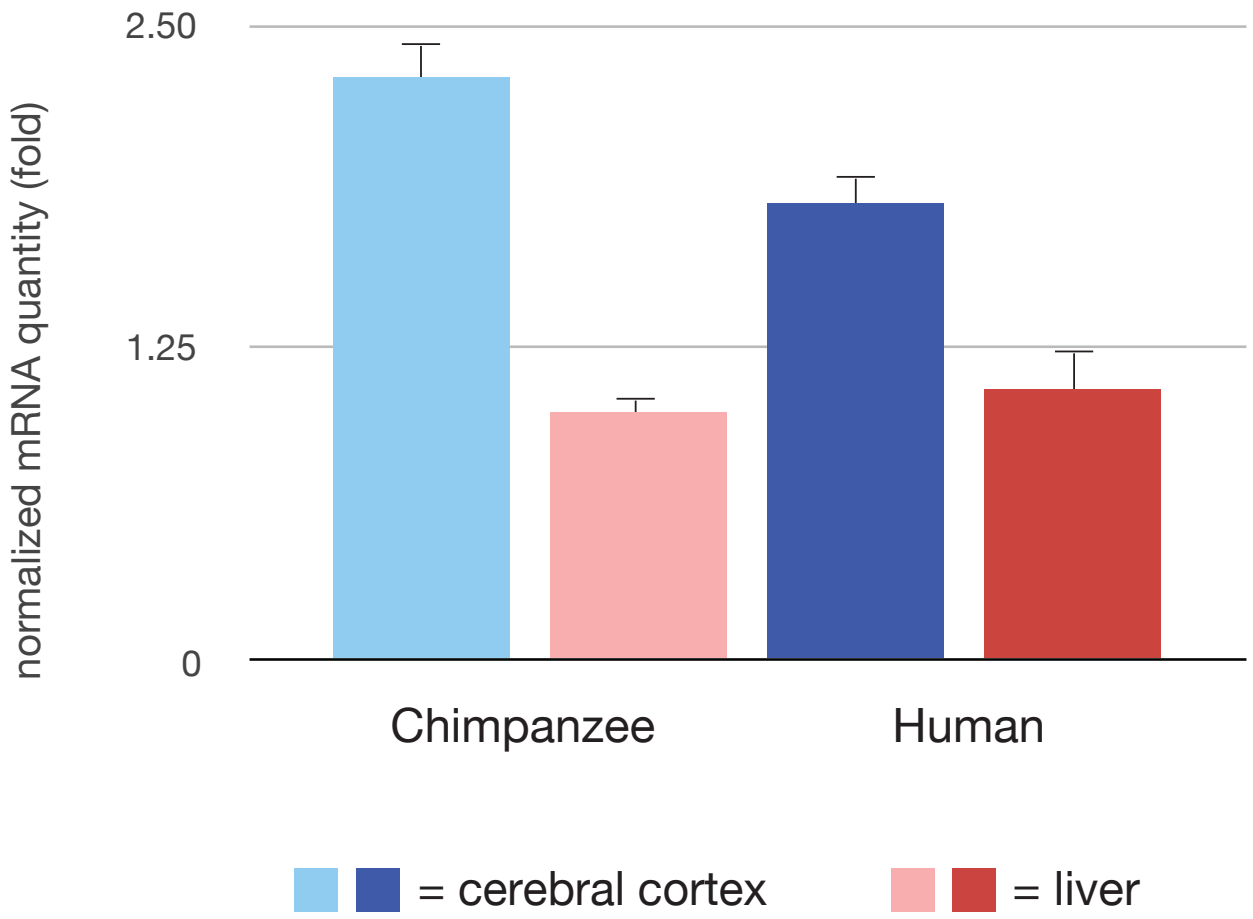
p-values; random-effects likelihood test for selection on 5 kb of 5' flanking region.

Chimpanzee 5’ flanking region drives higher expression



Dual-Luciferase assay. Tested in three different neuroblastoma cell lines.

Chimpanzees have higher *in vivo DRD4* expression in brain



Quantitative PCR on cDNA made from mRNA extracted from autopsied tissue.
Bars are means of six technical replicates.

What traits are affected?

Bonobo

food-rich, less seasonal habitat

abundant, nutritious back-up food

casual foragers, indifferent to food

risk-averse on standard psych profile

Chimpanzee

seasonal habitats

variable, nutrient poor back-up foods

active foragers, fixated on food

risk-tolerant on same profile

Part 3 summary: connecting the dots

	LCT	*DARC*	*DRD4*
genes of interest			
causal mutations	4 SNPs within distant module	2 SNPs within prox promoter	5 substitutions on chimp branch
molecular basis	increase tx level *in vivo* and *in vitro*; local epistasis	decrease tx level in erythrocytes *in vivo* and *in vitro*	increase tx level in cerebral cortex *in vivo* and *in vitro*
trait consequences	lactase persistence; dietary shift	resistance to parasite infection	cognition; foraging ecology?
eco and evo context	pastoralism; recent localized selection	parasite distribution; directional selection	ancient positive directional selection

Thanks to many people . . .

Computational

Ralph Haygood
Olivier Fédrigo
Jovana Pavisc
Brian Hanson
Ken Yokoyama

Humans

Courtney Babbitt
Matt Rockman
Jesse Silverman
Lisa Warner
William Nielsen
Tonya Severson
Adam Pfefferle
Shauna Morrow

Baboons

Jenny Tung
Alex Primus
Dagan Loisel
Andrew Bouley
Tonya Severson

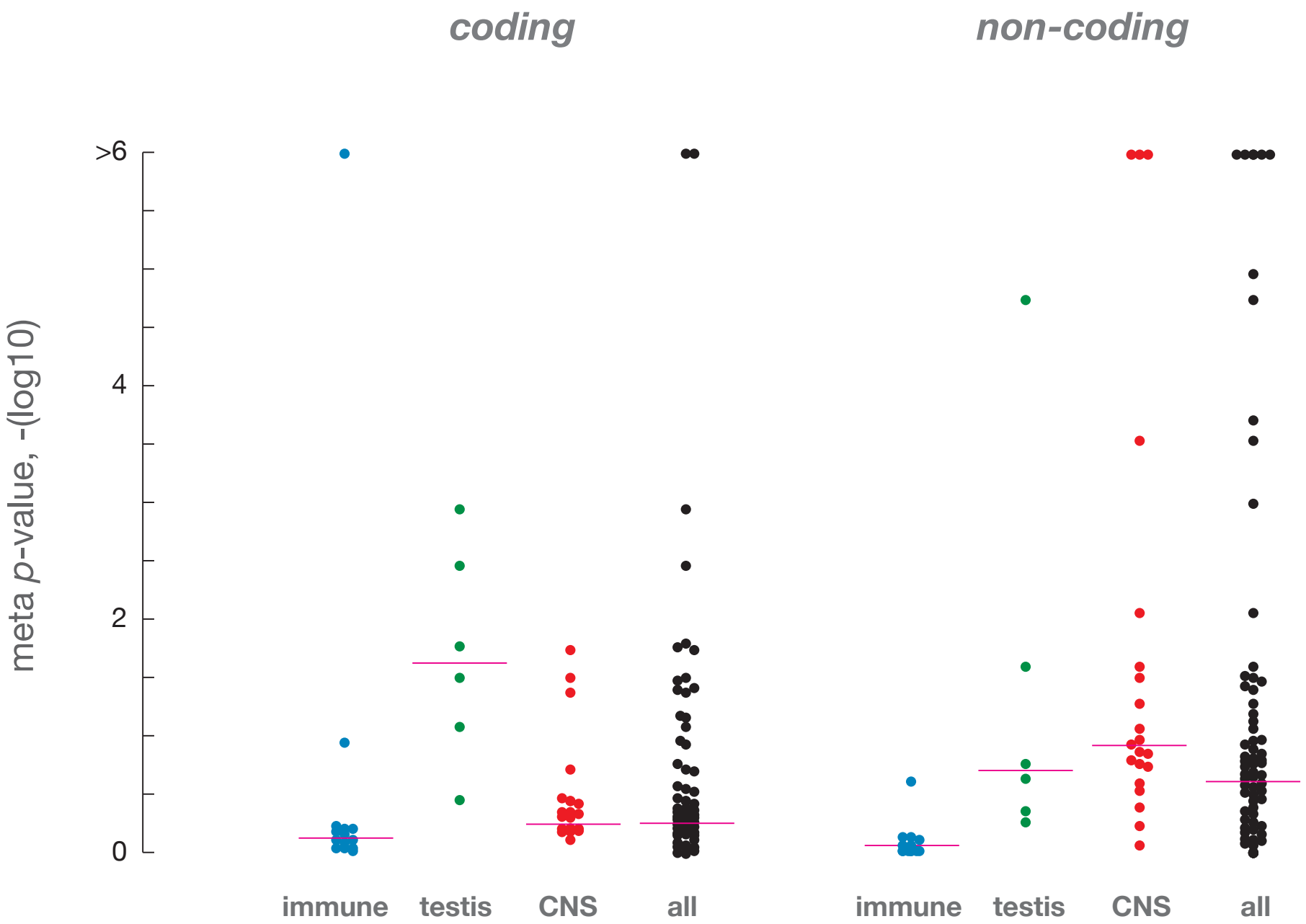
Help with programming

Jim Kent (UCSC)
Sergey Kosakovsy-Pond (UCSD)
Webb Miller (Penn)

Collaborators

Sarah Tishkoff (U MD)
Susan Alberts (Duke)
Jean Altmann (Princeton)

Qualitative differences in selection on coding/regulatory regions



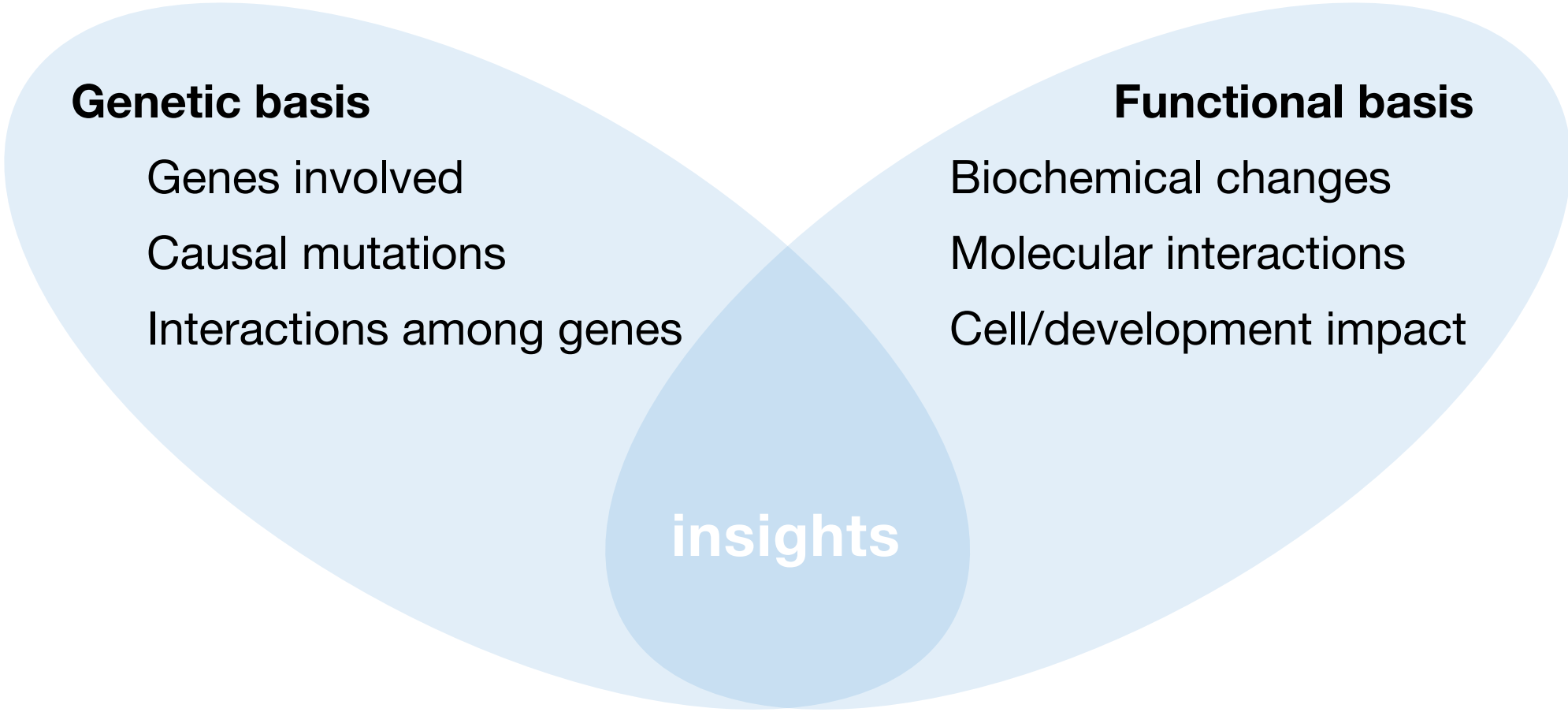
Understanding the evolution of gene expression

Genetic basis

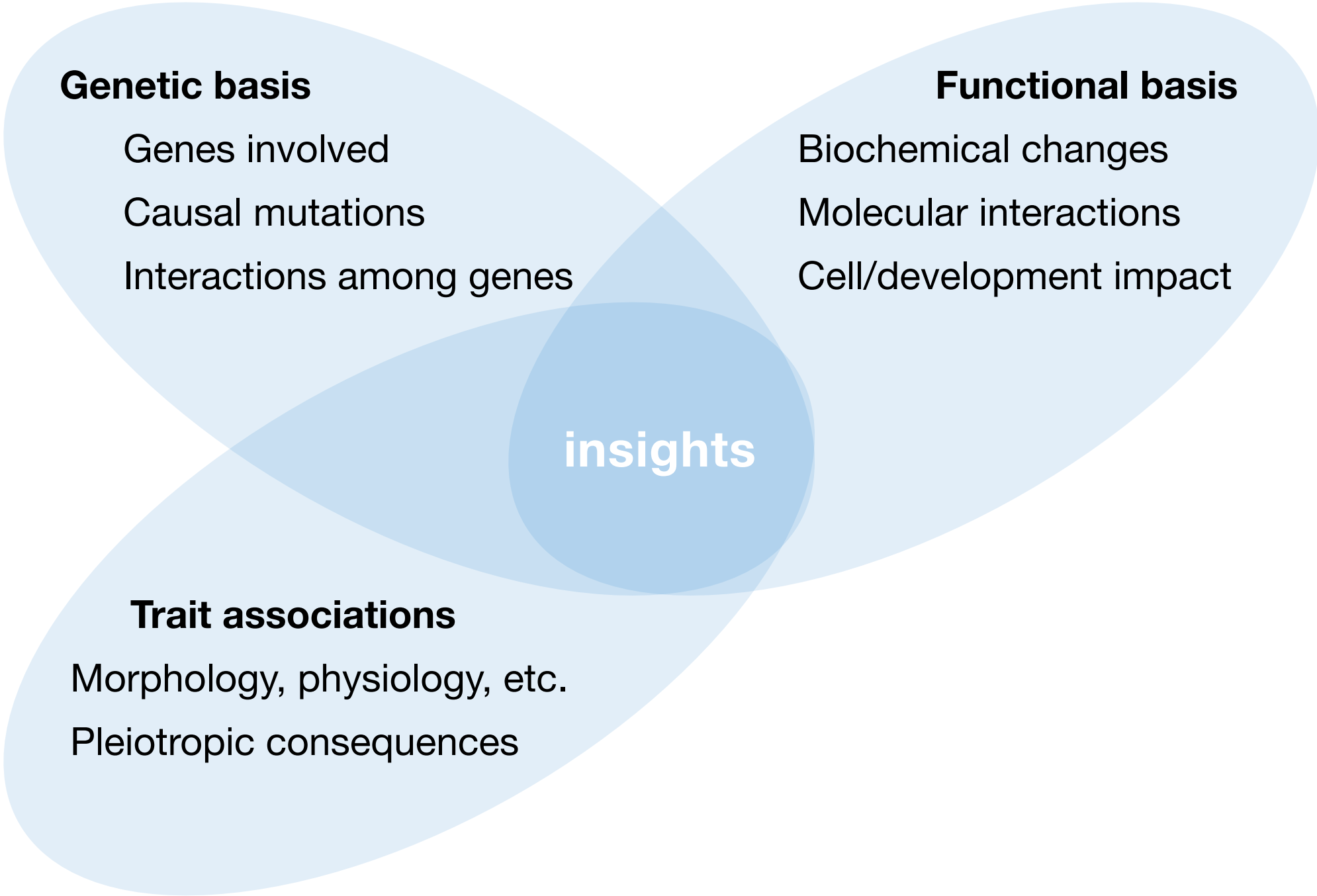
- Genes involved
- Causal mutations
- Interactions among genes

insights

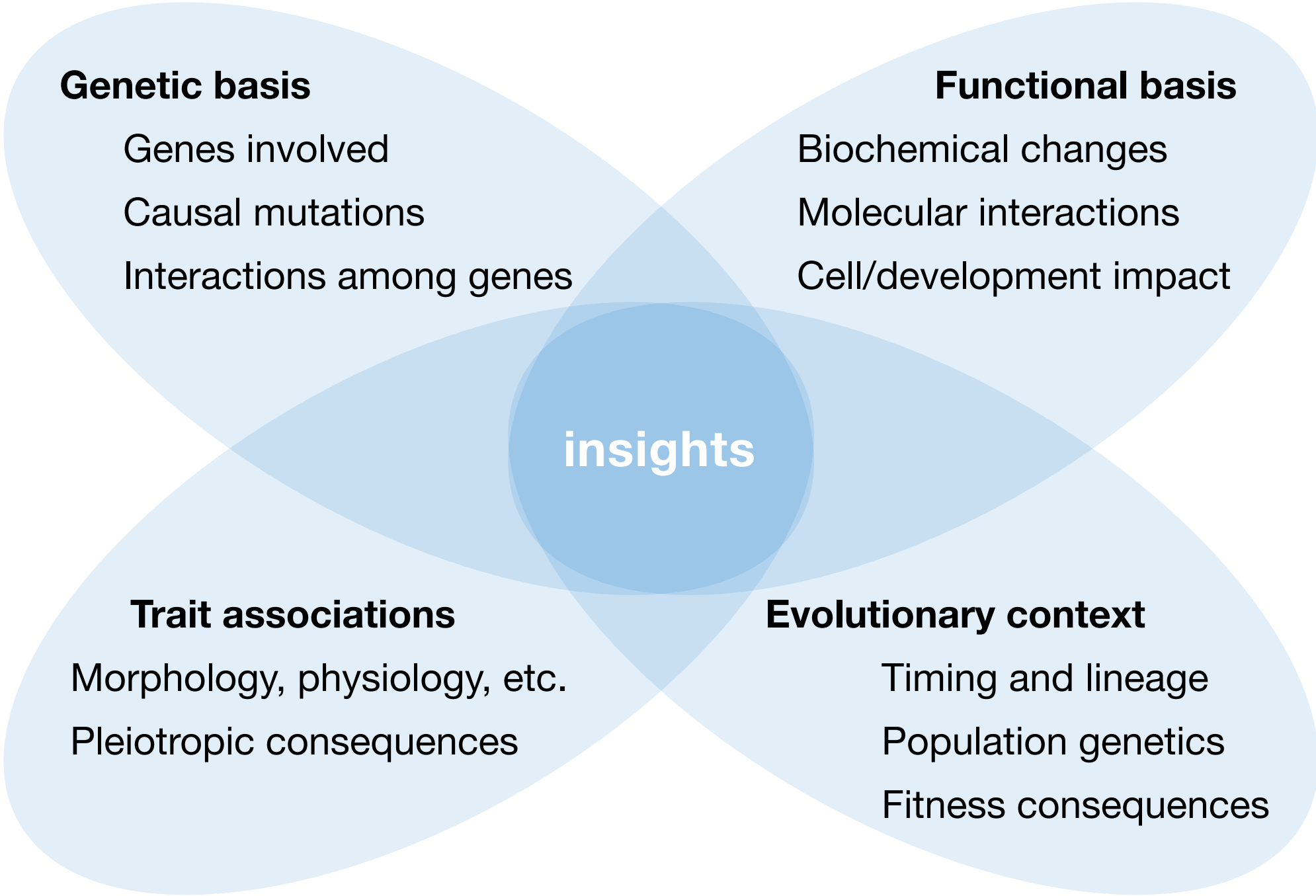
Understanding the evolution of gene expression



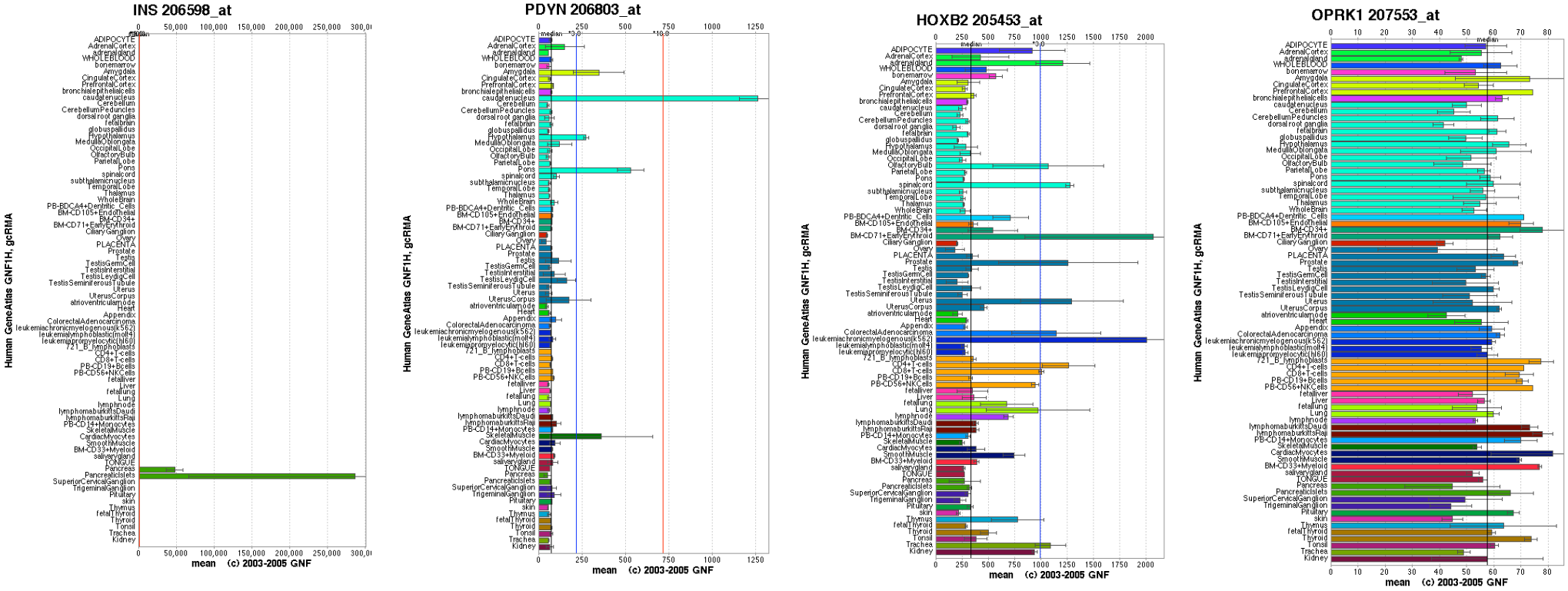
Understanding the evolution of gene expression



Understanding the evolution of gene expression



A metric for cell or tissue expression specificity



Data set: 73 non-cancerous tissues in the Novartis Gene Expression Atlas

Regard expression level of each tissue as a vector in 73-dimensional space

Specificity for each tissue = square of the cosine of the angle of the vector to that axis

Features

Independent of expression level

Can be defined for any gene in relation to any particular tissue